

WHOLE-GENOME PATTERNS OF DNA VARIATION IN MAIZE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Michael Allen Gore

May 2009

© 2009 Michael Allen Gore

WHOLE-GENOME PATTERNS OF DNA VARIATION IN MAIZE

Michael Allen Gore, Ph. D.

Cornell University 2009

Genome-Wide Association Studies (GWAS) offer the potential to resolve complex traits to a single gene or an individual polymorphism. GWAS, which rely on historical recombination for resolving complex traits, require that regions of active recombination be genotyped at high-density. Therefore, the molecular focus to the maize (*Zea mays* L.) genome is to genotype the gene-rich or low-copy-number regions, as these are the preferential sites for meiotic recombination. Due to the rapid rate of linkage disequilibrium decay in a large, diverse genome, it is possible that several million markers are needed for GWAS in diverse maize. The integration of gene-enrichment approaches and high-throughput genotyping platforms offer the potential to score polymorphisms at the needed scale in an efficient and cost-effective manner.

The two initial studies focused on developing methodologies to identify and score polymorphisms in large, complex plant genomes. In the first study, four gene-enrichment and complexity reduction target preparation methods were tested for scoring polymorphisms on the Affymetrix Maize GeneChip. The results indicated that the tested target preparation methods offered only modest power to detect polymorphisms with the Maize GeneChip. However, 10,000s of informative markers were still discovered. In the second study, gene-enriched genomic libraries constructed for two maize inbred lines were sequenced using massively parallel pyrosequencing. This combined with a computational SNP calling pipeline designed to reduce the

number of false positive SNPs resulting from paralogs lead to the identification of more than 120,000 single-nucleotide polymorphisms (SNPs).

The third study used Solexa sequencing for low-copy-enrichment resequencing of inbred lines that are the founders of the maize Nested Association Mapping (NAM) population. More than 3 million polymorphisms were scored across the founders, and a substantial portion of the low-copy fraction was highly divergent or novel relative to the reference genome. Recent and ancestral recombination rates were strongly correlated with nucleotide diversity, which suggests that genome structure partly shaped diversity. In addition, we identified regions of the maize genome that are potentially selective sweeps or involved in regional adaptation. These results should be an excellent resource for GWAS, fine-mapping projects, and understanding maize diversity and evolution.

BIOGRAPHICAL SKETCH

Michael Allen Gore was born December 22, 1975 in Fairfax, Virginia to Mr. Wayne Gore and Mrs. Carmen Rivera Gore. His grandfather Mr. Frederick Gore first introduced him to agriculture with countless hours of baling hay and mending fences. With his family's growing interest in horses, Michael moved to rural Culpeper, Virginia to raise Morgan horses. It is during this time when Michael developed a deep appreciation for agronomy and yearned to better understand the agronomic practices necessary to maximize crop productivity. This scholarly thirst was the impetus that led him to pursue a college degree at Virginia Polytechnic Institute and State University (Virginia Tech) in Blacksburg, Virginia.

Michael received his B.S. in Crop and Soil Environmental Sciences from Virginia Tech in December 1997 with a concentration in biotechnology and minors in chemistry and biology. While at Virginia Tech he conducted an undergraduate research project on the application of molecular markers to *Glycine max* (soybean) with Dr. M. A. Saghai Maroof. Upon successful completion of this project, he committed to further his academic career under the guidance of Dr. M. A. Saghai Maroof. In August 2000, he received a Master's degree from the Department of Crop and Soil Environmental Sciences with a specialization in crop molecular genetics and concentration in Molecular Cell Biology and Biotechnology (MCBB). His thesis project involved constructing a high-resolution genetic linkage map around the soybean virus resistance genes, *Rsv1* and *Rpv1*.

Upon completing his Master's degree, Michael decided to work in private industry. As a member of Rohm and Haas Company's plant gene regulation group in Spring House, Pennsylvania, he improved the design and efficacy of a chemical-inducible gene expression system for plants. After working at Rohm and Haas Company for a year, he relocated to Pioneer Hi-Bred International, Inc. in Johnston,

Iowa and contributed to the development of a maize hybrid breeding strategy that coupled SNP genotyping technologies and traditional plant breeding methods. Michael left Pioneer after two years of service and took an appointment as a contractor with Lancaster Laboratories, Inc for 10 months. In this new capacity, he applied comparative and functional genomics approaches to dissect biochemical pathways in tobacco at Philip Morris USA's genomics laboratory located in Richmond, Virginia.

After being immersed in private industry for a few years, Michael realized that a Ph.D. was essential for achieving his career goals and continuing job satisfaction. Therefore, he decided to reenter academia in 2004 and earn an advanced degree in Plant Breeding from Cornell University under the guidance of Dr. Edward S. Buckler. Michael decided to pursue a Ph.D. in plant breeding because of his long-term interest in understanding how to manipulate and preserve crop genetic diversity for sustainable agriculture systems. Michael has enjoyed his time as a Ph.D. student in the Department of Plant Breeding and Genetics and in Ed Buckler's group and knows that it was the right decision to pursue a Ph.D. at Cornell University.

I dedicate this dissertation to my patient, loving, and supportive wife Melanie,
and my precious daughters Ramona, Sasha, and Lucia.

ACKNOWLEDGMENTS

I would like to sincerely thank my advisor **Dr. Edward Buckler IV** for accepting me as a graduate student and giving me the freedom to pursue my research interests. Dr. Buckler has been an exceptional advisor throughout my graduate studies at Cornell University. I have greatly benefited from his wisdom, guidance, patience, and friendship. I have most enjoyed the many hours spent informally talking with him about plant genetics and diversity, because it is during these times that my passion for genetics and natural variation was nurtured as well as the refinement of my ability to ask important biological questions. Most importantly, I have learned from him how to balance family and work life. I am forever grateful to Dr. Buckler.

Also, I would like to thank my committee members **Drs. Stephen Kresovich** and **Timothy Setter**. In September 2003, Dr. Kresovich graciously hosted me for an impromptu visit to Ithaca and spent the afternoon showing me around campus. I have greatly benefited from his keen insights and constant encouragement throughout my time at Cornell University. Dr. Setter introduced me to the biology of how a plant responds to environmental stresses. It is this introduction that sparked my interests in plant physiology and biochemistry. I greatly appreciated all the times Dr. Setter would stay late after class to answer all of my intricate questions about plant physiology. I am a better scientist from having Drs. Kresovich and Setter on my committee.

I would like to thank all the past and present members of the Buckler lab that have contributed to my dissertation. **Dr. Elhan Ersoz** has assisted me tremendously throughout my research. She always shared with me her support, candor, and molecular biology skills. **Dr. Denise Costich** provided expert guidance on scientific writing and made the office a pleasant place to work. **Dr. Gael Pressoir** taught me European eating habits and how to apply quantitative genetics theory to plant

breeding. **Dr. Jianming Yu** introduced me to Henderson's mixed model for which I now have a deep appreciation. **Dr. Zhiwu Zhang** helped to expand my knowledge of statistics and challenged me to design and evaluate new algorithms. **Dallas Kroon** provided good humor and amazing computational skills to collate the massive amounts of phenotypic data. **Dr. Carlos Harjes** trained me in the field and was my inspiration to pursue the study of provitamin A and vitamin E in maize kernels. **Heather Yates** introduced me to the high-throughput candidate gene pipeline and showed me how to optimize PCR for diverse maize. **Dr. Feng Tian** provided great feedback on my research hypotheses and helped to solidify my understanding of positional cloning. **Terry Casstevens** helped to design the infrastructure and tools that managed and accessed the phenotypic and molecular data and was very supportive throughout my Ph.D. career. **Dr. Nengyi Zhang** helped me to better understand statistics and plant biochemistry. **Nick Lepak** was the field general and ensured that the plants were healthy and happy. Nick was always quick to provide assistance for my field and greenhouse experiments, but most importantly I valued his friendship. **Allison Krill** helped me with greenhouse and lab experiments, had a great sense of humor, and played good music in the lab. **Dr. Sean Myles** taught me about human genetics and was always willing to edit my papers or enthusiastically talk about science. **Dr. Amit Gur** introduced me to the power of metabolite profiling, enzyme activity assays, and how to carefully collect tissue samples. **Dr. Irie Vroh** provided me training in high-throughput sequencing and helped me to better understand the candidate gene selection process. **Dr. Matias Kirst** instructed me in the design and analysis of microarray studies and provided DNA samples for an array genotyping study. **Sara Larsson** and **Jason Peiffer** helped me with performing lab experiments and collecting immature ears. **Rob Elshire** constructed libraries and generated Solexa sequence data. The maize HapMap would never have been a reality if it was not for

Rob's dedication and hard work. **Huihui Li** showed me how to statistically analyze genetic data and kept me entertained with her wit. **Dr. Peter Bradbury** was a tremendous collaborator. He was always prompt with providing results and text for manuscripts, and patiently showed me how to analyze data with mixed models in SAS. **Kevin Howe** helped with providing resources for my research and introduced me to the best food in Ithaca. **Natalie Stevens** and **Linda Rigamer Lirette** provided excellent technical editing of manuscripts. **Emily Briggs, Emily Gordon, Gregori Temnykh,** and **Michelle Denton** were undergraduate researchers that contributed to my lab and field experiments. They provided tremendous help with my lab, greenhouse, and field experiments, and I have no doubt that they will go on to do great things.

The Institute of Genomic Diversity has provided physical and intellectual resources to support my research. In particular, I would like to thank **Wen Yen Zhu, Charlotte Acharya, Hong Sun,** and **Drs. Sharon Mitchell, Alexandra Casa, Martha Hamblin, Patrick Brown,** and **Seth Murray.**

My collaborators at Keygene N.V., **René Hogers, Esther Verstege, Jan van Oeveren, Johan Peleman,** and **Michiel van Eijk,** were instrumental in designing the array genotyping experiment and generating high quality data. Also, my collaborators at 454 Life Sciences, **Pascal Bouffard, Edward Szekeres, Thomas Jarvie,** and Roche Applied Science, Corporation, **Timothy Harkins,** generated the pyrosequencing data used to call SNPs between maize inbred lines B73 and Mo17. **Mark Wright** is a brilliant computational biologist and population geneticist that helped to analyze the pyrosequencing data and developed the method to prevent false positive SNP calls from paralogous sequence alignments. My collaborators at Cold Spring Harbor Laboratory, **Dr. Doreen Ware, Bonnie Hurwitz, Apurva Narechania,** and **Jer-ming Chia,** are talented computational biologists that helped to

analyze the pyrosequencing and Solexa sequencing data. **Dr. George Grills**, the Director of the Life Sciences Core Laboratories Center at Cornell University, helped to establish collaborations with 454 Life Sciences, Roche Applied Science, Corporation, and Illumina. If it was not for **Dr. Grills**, the Solexa sequencing instrument might still be broken. The other investigators on the Maize Diversity Project, **Drs. John Doebley, Brandon Gaut, Major Goodman, James Holland, Michael McMullen, Lincoln Stein**, and their respective research groups generated materials, germplasm, resources and data that helped to improve my research.

Cornell University is an exciting place to conduct research in the plant sciences. The faculty, staff, and students in the **Department of Plant Breeding and Genetics** were vital to making my Ph.D. experience at Cornell University an amazing experience.

I wish to thank the **Department of Plant Breeding and Genetics**, the **United States Department of Agriculture**, and the **National Science Foundation** (DBI-0321467 and DBI-0638566) for support of my graduate studies and research projects.

I would like to thank all of my supportive friends and colleagues I met while at Cornell University. Most importantly, I would like to thank my family for their support and understanding.

TABLE OF CONTENTS

BIOGRAPHICAL SKETCH	iii
DEDICATION	v
ACKNOWLEDGMENTS	vi
TABLE OF CONTENTS	x
LIST OF TABLES	xi
LIST OF FIGURES	xii
 CHAPTER 1 GENOTYPING FOR ASSOCIATION MAPPING IN PLANTS	 1
REFERENCES	9
 CHAPTER 2 EVALUATION OF TARGET PREPARATION METHODS FOR SINGLE FEATURE POLYMORPHISM DETECTION IN LARGE COMPLEX PLANT GENOMES	 14
ABSTRACT	14
INTRODUCTION	15
MATERIALS AND METHODS	18
RESULTS	28
DISCUSSION	38
REFERENCES	45
 CHAPTER 3 LARGE-SCALE ENRICHMENT AND DISCOVERY OF GENE- ENRICHED SNPS	 50
ABSTRACT	50
INTRODUCTION	51
MATERIALS AND METHODS	53
RESULTS	62
DISCUSSION	77
REFERENCES	84
 CHAPTER 4 A FIRST GENERATION HAPLOTYPE MAP OF THE WORLD'S MOST DIVERSE CROP: MAIZE	 90
ABSTRACT	90
INTRODUCTION	91
MATERIALS AND METHODS	92
RESULTS AND DISCUSSION	101
REFERENCES	118

LIST OF TABLES

Table 2.1. SFP detection potential of target preparation methods.	32
Table 2.2. Mixed model analysis of SFP detection power.	33
Table 2.3. Mixed model analysis of SFP detection power with shared probes.	35
Table 2.4. The distribution of SNP position in probes that detect a single SNP.	36
Table 2.5. Mixed model analysis of polymorphic probeset detection power.	38
Table 2.6. Estimated and observed number of true SFPs that were identified on the whole Maize GeneChip.	41
Table 3.1. Statistics of databases and genome sequences used in this study.	58
Table 3.2. Sequence composition of modified HMPR and UF libraries.	65
Table 3.3 Gene Enrichment Analysis of modified HMPR and UF libraries.	68
Table 3.4. Summary of the assembly process.	70
Table 3.5. Summary of putative SNPs and call rates at various coverage depths and quality value thresholds with and without implementation of the paralog distinguishing list (PDL).	74
Table 3.6. Summary of B73/Mo17 454 SNP validation.	76

LIST OF FIGURES

Figure 1.1. Comparison of sequencing platforms for high-throughput SNP Discovery.	6
Figure 2.1. Frequency distribution of perfect match (PM) and mismatch (MM) probe pair signal intensity ratios.	30
Figure 3.1. Illustration of a recent single gene duplication event that results in highly similar paralogs, and how the paralog distinguishing list (PDL) distinguishes alleles from paralogs when calling SNPs.	71
Figure 4.1. Flowchart of how methylation-filtration (MF) <i>HpaII</i> , whole-genome amplification (WGA) <i>HpaII</i> , and <i>BbvI</i> genomic libraries were constructed.	94
Figure 4.2. Percentage (%) of low-copy, high-copy, and unmapped SBS reads by library type.	103
Figure 4.3. Number of Megabases (Mb) in the maize genome that was sequenced versus depth of genome base coverage.	104
Figure 4.4. Distribution of BAC inter-marker distances for 2,867,766 non-redundant SNPs and indels.	105
Figure 4.5. Distribution of unmapped low-copy-number SBS reads across the founder lines.	107
Figure 4.6. Minor allele frequency (MAF) distributions for SNPs discovered by Sanger and Illumina sequencing technologies on 27 diverse maize lines.	108
Figure 4.7. Autocorrelation plot of pairwise nucleotide diversity (π) and population-recombination (C) as a function of physical distance.	109
Figure 4.8. Genome-wide patterns of pairwise nucleotide diversity (π) and population-recombination (C).	110
Figure 4.9. Correlations between two estimates of recombination (R and C) and between estimates of recombination and nucleotide diversity (π).	113
Figure 4.10. Plot of the correlation between sequence features and two estimates of recombination (R and C) as well as pairwise nucleotide diversity (π).	114
Figure 4.11. The average decay of LD across the length of a BAC.	117

CHAPTER 1

GENOTYPING FOR ASSOCIATION MAPPING IN PLANTS¹

Background Markers

In association studies, a set of unlinked, selectively neutral background markers scaled to achieve genome-wide coverage are employed to broadly characterize the genetic composition of individuals. Background genetic markers are useful in assigning individuals to populations (Pritchard and Rosenberg, 1999), preventing spurious associations if population structure and relatedness exist (Pritchard et al., 2000; Thornsberry et al., 2001; Yu et al., 2006), and estimating kinship and inbreeding (Lynch and Ritland, 1999). Random amplified polymorphic DNA (RAPD) (Williams et al., 1990) and amplified fragment length polymorphism (AFLP) (Vos et al., 1995) markers can serve as background markers, but almost all RAPD and AFLP markers are dominantly inherited and thus demand special statistical methods if used to estimate population genetic parameters (Falush et al., 2007; Ritland, 2005). Conversely, codominant microsatellites, or simple sequence repeats (SSRs), and single-nucleotide polymorphisms (SNPs) are more revealing (i.e., no allelic ambiguity) than their dominant counterparts and, therefore, are more powerful in estimating population structure (Q) and the relative kinship matrix (K).

Because SSR markers are multiallelic, reproducible, PCR-based, and generally selectively neutral they have been the predominant molecular marker in kinship and population studies. Semi-automated systems exist for the multiplexed detection and sizing of fluorescent-labeled SSR products with internal size standards; thus greatly increasing both the allele size accuracy and genotyping throughput (Mitchell et al.,

¹ This introduction was published as part of a peer-reviewed review article. C. Zhu, M. Gore, E. Buckler, and J. Yu. 2008. Status and Prospects of Association Mapping in Plants. *Plant Genome* 1: 5-20.

1997). Nascent polymorphic SSR alleles are mostly spawned from the slipped strand mispairing (i.e., slippage) of allelic tandem repeats during DNA replication (Levinson and Gutman, 1987). In theory, the highly mutagenic process of slippage can generate an unlimited number of SSR alleles, but longer SSR allele sizes are more likely to be eliminated by natural selection (Li et al., 2002). The same slippage phenomenon that results in highly polymorphic SSR loci also is the basis of size homoplasy, a situation when SSR alleles are identical in size but not identical by descent (Viard et al., 1998). If alleles have a high mutation rate and strong size constraint, SSR size homoplasy could be problematic when estimating genetic parameters in a large population (Estoup et al., 2002).

Due to higher genome density, lower mutation rate, and better amenability to high-throughput detection systems, SNPs are rapidly becoming the marker of choice for complex trait dissection studies. Either single marker assays or multiplexes in scalable assay plates and microarray formats can be used to score SNPs. The selection of a specific genotyping technology is dependent on both the number of SNP markers and individuals to be scored (Kwok, 2000; Syvänen, 2005). The mutation rate per site per generation is several times lower than the SSR mutational rate per generation (Li et al., 2002; Vigouroux et al., 2002). Therefore, on a per-site basis, due to SNPs' predominantly biallelic nature they are less informative than multiallelic SSRs. Because the expected heterozygosity of individual SNPs is lower, more SNP than SSR background markers are needed to reach a reasonable estimate of population structure and relatedness for most crops (Hamblin et al., 2007). This should not be considered a shortcoming because SNPs are more widely distributed throughout the genome and are several-fold less expensive to score than SSRs.

Candidate Genes

Candidate-gene association mapping is a hypothesis-driven approach to complex trait dissection, with biologically relevant candidates selected and ranked based on the evaluation of available results from genetic, biochemical, or physiology studies in model and non-model plant species (Mackay, 2001; Risch and Merikangas, 1996). Because SNPs offer the highest resolution for mapping QTL and are potentially in LD with the causative polymorphism they are the preferential candidate-gene variant to genotype in association studies (Rafalski, 2002). Candidate-gene association mapping requires the identification of SNPs between lines and within specific genes. Therefore, the most straightforward method of identifying candidate gene SNPs relies on the resequencing of amplicons from several genetically distinct individuals of a larger association population. Fewer diverse individuals in the SNP discovery panel are needed to identify common SNPs, whereas many more are needed to identify rarer SNPs. Promoter, intron, exon, and 5'/3'-untranslated regions are all reasonable targets for identifying candidate gene SNPs, with non-coding regions expected to have higher levels of nucleotide diversity than coding regions. The rate of LD decay for a specific candidate gene locus dictates the number of SNPs per unit length (e.g., kb) needed to identify significant associations (Whitt and Buckler, 2003). Therefore, the number and base-pair length of amplicons required to sufficiently sample a candidate gene locus is almost entirely dependent on LD and SNP distribution, with a higher density of SNP markers needed in regions of relatively low LD and high nucleotide diversity.

It is not essential to score every candidate gene SNP. Because a key objective of this approach is to identify SNPs that are causal of phenotypic variation, those with a higher likelihood to alter protein function (coding SNPs) or gene expression (regulatory SNPs) should be a top priority for genotyping (Tabor et al., 2002). However, the biological function of SNPs, if any, for the most part is unknown or not

easily discerned. In cases of ambiguity where there are blocks of several SNPs in significant LD, an alternative strategy is to select and score a small fraction of SNPs (tag SNPs) that capture most of the haplotype block structure in candidate-gene regions (Johnson et al., 2001). Genotyping tag SNPs is more cost effective and, if properly designed, does not result in a significant loss of statistical testing power (Kui et al., 2002). In most cases, allele resequencing in diploid inbred lines (homozygous loci) allows for the direct determination of haplotypes. Reconstructing haplotypes from SNP data in heterozygous and polyploid (ancient or modern) individuals is more challenging, as statistical algorithms are needed to resolve phase ambiguities (Simko, 2004; Stephens et al., 2001) and transmission tests are needed to confirm orthologous relationships (Cogan et al., 2007).

Candidate-gene selection is straightforward for relatively simple biochemical pathways (e.g., starch synthesis in maize) or well characterized pathways (e.g., flowering time in *Arabidopsis*) that have been resolved mainly through genetic analysis of mutant loci (natural or induced). But for complex traits such as grain or biomass yield, the entire genome could potentially serve as a candidate (Yu and Buckler, 2006). Most candidate-gene studies investigating a single pathway or trait in a crop species have genotyped less than 100 SNPs in a population of 100 to 400 individuals (Ersoz et al., 2008). In these studies, Sanger sequencing and single base extension (SBE) assays were the predominant technologies used to score candidate gene SNPs. Advantages of SBE assays over Sanger sequencing are reflected in their lower reagent costs, enhanced resolution of heterozygous genotypes, and better suitability to multiplex detection on higher-throughput, lower cost analytical platforms (Syvänen, 2001).

Whole-Genome Scan

If whole-genome association scans are to be conducted in crops, an important first step is to use high-capacity DNA sequencing instruments or high-density oligonucleotide (oligo) arrays to efficiently identify SNPs at a density that accurately reflects genome-wide LD structure and haplotype diversity. The appropriateness of a DNA sequencing platform (Figure 1.1) for SNP discovery depends on the number of SNPs required for effective whole-genome scans in an association population. For example, the extensive LD in 95 *Arabidopsis* accessions and 102 elite barley inbred lines made it possible to association test a low number of evenly spaced SNPs discovered via capillary-based Sanger sequencing and still achieve a medium level of genome-wide mapping resolution (Aranzana et al., 2005; Rostoks et al., 2006). Alternatively, tens to hundreds of thousands of SNP markers are required for powerful whole-genome scans in crops with low LD and high haplotype diversity, such as maize and sunflower. In such a scenario, the 454-GS FLX (Margulies et al., 2005) and Illumina 1 G Genome Analyzer (Bennett, 2004) are ideal platforms for identifying scores of SNPs through short read resequencing of allelic fragments from several genetically diverse individuals. After SNPs are identified, different array-based platforms can be used to genotype thousands of tag SNPs in parallel.

A high quality whole-genome reference sequence is extremely valuable in construction of a SNP haplotype map from short reads produced by the 454 and Illumina sequencing platforms. This is because short reads are more easily assembled by aligning to a preexisting genome reference sequence compared to de novo assembly. Also, a reference genome is useful in masking repetitive and paralogous sequences, as the orthology of high copy sequences is difficult to determine unless candidate SNPs are genetically mapped. Because the base calling accuracy of 454 and Illumina is presently lower than that of Sanger sequencing, emphasis should be placed

on calling SNPs that have multiple read support ($\geq 2X$ coverage/allele/individual). The newness and expense of next-generation sequencing technologies have limited their wide-spread implementation for SNP discovery in crops. Recently, a 454-based transcriptome sequencing method was used in maize to identify more than 36,000 candidate SNPs between two maize inbred lines (Barbazuk et al., 2007). This 454-SNP study is a promising step toward development of numerous genome-wide SNP markers in a highly diverse crop species with a rapid breakdown of LD, but more importantly lays the framework for identifying SNPs based on sequencing of random genomic fragments.

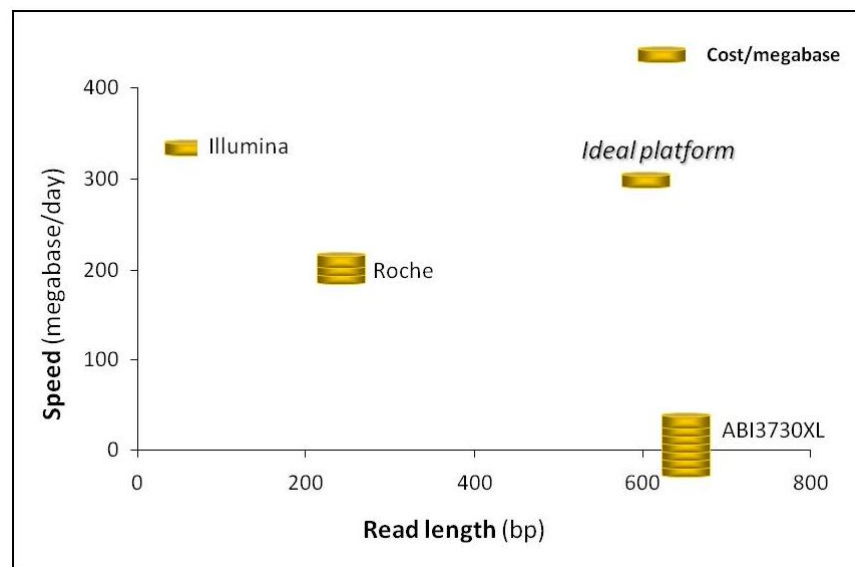


Figure 1.1 Comparison of sequencing platforms for high-throughput SNP discovery.

The simultaneous discovery and genotyping of allelic variation with high-density oligo expression arrays designed from a reference sequence is based on the concept that a perfectly matched target binds to a 25-bp oligo feature with greater affinity than a mismatched target (Borevitz et al., 2003; Winzeler et al., 1998). If an

individual feature on an array shows a significant and repeatable difference in hybridization intensity between genotypes, it can serve directly as a polymorphic marker or single feature polymorphism (SFP). Expression arrays hybridized with total genomic DNA allow for highly accurate scoring of several thousand SFPs in the relatively small genomes of 135-Mb *Arabidopsis* (Borevitz et al., 2003) and 430-Mb rice (Kumar et al., 2007). Whole-genome, genome complexity reduction, and gene enrichment target preparation methods are only modestly successful for detecting SFPs in larger retrotransposon-rich plant genomes (Gore et al., 2007; Rostoks et al., 2005). Notable limitations are that SFPs tend to be less heritable (i.e., lower quality) than SNPs and map unknown polymorphisms only at 25-bp resolution. If scored at very high density and moderate accuracy, SFPs are potentially powerful tools to detect associations in crop genomes with extensive LD (Kim et al., 2006) and relatively low levels of repetitive DNA.

In a whole-genome resequencing-by-hybridization approach championed by Perlegen Sciences (Mountain View, CA), high-density arrays consisting of tiled, overlapping 25-bp oligos are used to identify SNPs and other polymorphisms in a hybridized target genome at single base pair resolution (Borevitz and Ecker, 2004; Mockler et al., 2005). Tiling arrays were used to construct a haplotype map by essentially resequencing 20 diverse *Arabidopsis* genomes and cataloging more than 1 million nonredundant SNPs (Clark et al., 2007). Only 27% of the total polymorphisms were scored in a given ecotype due to ineffective SNP detection in highly polymorphic regions. Tiling array projects are in progress to identify SNPs in multiple rice lines (McNally et al., 2006) and score 250,000 tag SNPs in an association panel of 1,000 *Arabidopsis* ecotypes. It is still an open question as to whether resequencing-by-hybridization on tiling arrays will come to fruition as a routine SNP discovery

platform for crop genomes that predominantly contain repetitive DNA, extensive sequence duplications, or high nucleotide diversity.

REFERENCES

- Aranzana, M.J., S. Kim, K. Zhao, E. Bakker, M. Horton, K. Jakob, C. Lister, J. Molitor, C. Shindo, C. Tang, C. Toomajian, B. Traw, H. Zheng, J. Bergelson, C. Dean, P. Marjoram, and M. Nordborg. 2005. Genome-wide association mapping in *Arabidopsis* identifies previously known flowering time and pathogen resistance genes. *PLoS Genet* 1:e60.
- Barbazuk, W.B., S.J. Emrich, H.D. Chen, L. Li, and P.S. Schnable. 2007. SNP discovery via 454 transcriptome sequencing. *Plant J.* 51:910-918.
- Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics* 5:433-438.
- Borevitz, J.O., and J.R. Ecker. 2004. Plant genomics: the third wave. *Annu. Rev. Genomics Hum. Genet.* 5:443-477.
- Borevitz, J.O., D. Liang, D. Plouffe, H.S. Chang, T. Zhu, D. Weigel, C.C. Berry, E. Winzeler, and J. Chory. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 13:513-523.
- Clark, R.M., G. Schweikert, C. Toomajian, S. Ossowski, G. Zeller, P. Shinn, N. Warthmann, T.T. Hu, G. Fu, D.A. Hinds, H. Chen, K.A. Frazer, D.H. Huson, B. Scholkopf, M. Nordborg, G. Ratsch, J.R. Ecker, and D. Weigel. 2007. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338-342.
- Cogan, N.O., M.C. Drayton, R.C. Ponting, A.C. Vecchies, N.R. Bannan, T.I. Sawbridge, K.F. Smith, G.C. Spangenberg, and J.W. Forster. 2007. Validation of in silico-predicted genic SNPs in white clover (*Trifolium repens* L.), an outbreeding allopolyploid species. *Mol. Genet. Genomics* 277:413-425.
- Ersoz, E., J. Yu, and E.S. Buckler. 2008. Applications of linkage disequilibrium and association mapping in crop plants, p. 97-120, *In* R. Varshney and R. Tuberosa, eds. *Genomics-Assisted Crop Improvement, Vol. 1 Genomics Approaches and Platforms*. Springer Verlag.
- Estoup, A., P. Jarne, and J.-M. Cornuet. 2002. Homoplasmy and mutation model at microsatellite loci and their consequences for population genetics analysis. *Mol. Ecol.* 11:1591-1604.
- Falush, D., M. Stephens, and J.K. Pritchard. 2007. Inference of population structure using multilocus genotype data: dominant markers and null alleles. *Mol. Ecol. Notes* 7:574-578.

- Gore, M., P. Bradbury, R. Hogers, M. Kirst, E. Verstege, J. van Oeveren, J. Peleman, E. Buckler, and M. van Eijk. 2007. Evaluation of Target Preparation Methods for Single-Feature Polymorphism Detection in Large Complex Plant Genomes. *Crop Sci.* 47:S-135-148.
- Hamblin, M.T., M.L. Warburton, and E.S. Buckler. 2007. Empirical Comparison of Simple Sequence Repeats and Single Nucleotide Polymorphisms in Assessment of Maize Diversity and Relatedness. *PLoS ONE* 2:e1367.
- Johnson, G.C.L., L. Esposito, B.J. Barratt, A.N. Smith, J. Heward, G. Di Genova, H. Ueda, H.J. Cordell, I.A. Eaves, F. Dudbridge, R.C.J. Twells, F. Payne, W. Hughes, S. Nutland, H. Stevens, P. Carr, E. Tuomilehto-Wolf, J. Tuomilehto, S.C.L. Gough, D.G. Clayton, and J.A. Todd. 2001. Haplotype tagging for the identification of common disease genes. *Nat. Genet.* 29:233-237.
- Kim, S., K. Zhao, R. Jiang, J. Molitor, J.O. Borevitz, M. Nordborg, and P. Marjoram. 2006. Association mapping with single-feature polymorphisms. *Genetics* 173:1125-1133.
- Kui, Z., P. Calabrese, M. Nordborg, and S. Fengzhu. 2002. Haplotype Block Structure and Its Applications to Association Studies: Power and Study Designs. *Am. J. Hum. Genet.* 71:1386-1394.
- Kumar, R., J. Qiu, T. Joshi, B. Valliyodan, D. Xu, and H.T. Nguyen. 2007. Single feature polymorphism discovery in rice. *PLoS ONE* 2:e284.
- Kwok, P.Y. 2000. High-throughput genotyping assay approaches. *Pharmacogenomics* 1:95-100.
- Levinson, G., and G.A. Gutman. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol. Biol. Evol.* 4:203-221.
- Li, Y.-C., A.B. Korol, T. Fahima, A. Beiles, and E. Nevo. 2002. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. *Mol. Ecol.* 11:2453-2465.
- Lynch, M., and K. Ritland. 1999. Estimation of pairwise relatedness with molecular markers. *Genetics* 152:1753-1766.
- Mackay, T.F. 2001. The genetic architecture of quantitative traits. *Annu. Rev. Genet.* 35:303-39.
- Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.-J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C.

- Jando, M.L.I. Alenquer, T.P. Jarvie, K.B. Jirage, J.-B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
- McNally, K.L., R. Bruskiewich, D. Mackill, C.R. Buell, J.E. Leach, and H. Leung. 2006. Sequencing Multiple and Diverse Rice Varieties. Connecting Whole-Genome Variation with Phenotypes. *Plant Physiol.* 141:26-31.
- Mitchell, S.E., S. Kresovich, C.A. Jester, C.J. Hernandez, and A.K. Szewc-McFadden. 1997. Application of multiplex PCR and fluorescence-based, semi-automated allele sizing technology for genotyping plant genetic resources. *Crop Sci.* 37:617-624.
- Mockler, T.C., S. Chan, A. Sundaresan, H. Chen, S.E. Jacobsen, and J.R. Ecker. 2005. Applications of DNA tiling arrays for whole-genome analysis. *Genomics* 85:1-15.
- Pritchard, J.K., and N.A. Rosenberg. 1999. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* 65:220-228.
- Pritchard, J.K., M. Stephens, N.A. Rosenberg, and P. Donnelly. 2000. Association mapping in structured populations. *Am. J. Hum. Genet.* 67:170-181.
- Rafalski, A. 2002. Applications of single nucleotide polymorphisms in crop genetics. *Curr. Opin. Plant Biol.* 5:94-100.
- Risch, N., and K. Merikangas. 1996. The future of genetic studies of complex human diseases. *Science* 273:1516-1517.
- Ritland, K. 2005. Multilocus estimation of pairwise relatedness with dominant markers. *Mol. Ecol.* 14:3157-3165.
- Rostoks, N., J.O. Borevitz, P.E. Hedley, J. Russell, S. Mudie, J. Morris, L. Cardle, D.F. Marshall, and R. Waugh. 2005. Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.* 6:R54.
- Rostoks, N., L. Ramsay, K. MacKenzie, L. Cardle, P.R. Bhat, M.L. Roose, J.T. Svensson, N. Stein, R.K. Varshney, D.F. Marshall, A. Graner, T.J. Close, and R. Waugh. 2006. Recent history of artificial outcrossing facilitates whole-

- genome association mapping in elite inbred crop varieties. *Proc. Natl. Acad. Sci.* 103:18656-18661.
- Simko, I. 2004. One potato, two potato: haplotype association mapping in autotetraploids. *Trends Plant Sci.* 9:441-448.
- Stephens, M., N.J. Smith, and P. Donnelly. 2001. A new statistical method for haplotype reconstruction from population data. *Am. J. Hum. Genet.* 68:978-989.
- Syvänen, A.C. 2001. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.* 2:930-942.
- Syvänen, A.C. 2005. Toward genome-wide SNP genotyping. *Nat. Genet.* 37 Suppl:S5-10.
- Tabor, H.K., N.J. Risch, and R.M. Myers. 2002. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat. Rev. Genet.* 3:391-397.
- Thornsberry, J.M., M.M. Goodman, J. Doebley, S. Kresovich, D. Nielsen, and E.S.t. Buckler. 2001. Dwarf8 polymorphisms associate with variation in flowering time. *Nat. Genet.* 28:286-289.
- Viard, F., P. Franck, M.-P. Dubois, A. Estoup, and P. Jarne. 1998. Variation of Microsatellite Size Homoplasmy Across Electromorphs, Loci, and Populations in Three Invertebrate Species. *J. Mol. Evol.* 47:42-51.
- Vigouroux, Y., J.S. Jaqueth, Y. Matsuoka, O.S. Smith, W.D. Beavis, J.S.C. Smith, and J. Doebley. 2002. Rate and Pattern of Mutation at Microsatellite Loci in Maize. *Mol. Biol. Evol.* 19:1251-1260.
- Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, and M. Kuiper. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23:4407-4414.
- Whitt, S.R., and E.S. Buckler. 2003. Using natural allelic diversity to evaluate gene function. *Methods Mol. Biol.* 236:123-140.
- Williams, J.G.K., A.R. Kubelik, K.J. Livak, J.A. Rafalski, and S.V. Tingey. 1990. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* 18:6531-6535.
- Winzeler, E.A., D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, and

- R.W. Davis. 1998. Direct allelic variation scanning of the yeast genome. *Science* 281:1194-1197.
- Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. *Curr. Opin. Biotech.* 17:155-160.
- Yu, J., G. Pressoir, W.H. Briggs, I. Vroh Bi, M. Yamasaki, J.F. Doebley, M.D. McMullen, B.S. Gaut, D.M. Nielsen, J.B. Holland, S. Kresovich, and E.S. Buckler. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat. Genet.* 38:203-208.

CHAPTER 2

EVALUATION OF TARGET PREPARATION METHODS FOR SINGLE FEATURE POLYMORPHISM DETECTION IN LARGE COMPLEX PLANT GENOMES²

ABSTRACT

For those genomes low in repetitive DNA, hybridizing total genomic DNA to high-density expression arrays offers an effective strategy for scoring single feature polymorphisms (SFPs). Of the ~2.5 Gb that constitute the maize genome (*Zea mays* L.), only 10-20% are genic sequences, with large amounts of repetitive DNA intermixed throughout. Therefore, a target preparation method engineered to generate a high genic-to-repetitive DNA ratio is essential for SFP detection in maize. To that end, we tested four gene enrichment and complexity reduction target preparation methods for scoring SFPs on the Affymetrix GeneChip Maize Genome Array (“Maize GeneChip”). Methylation filtration (MF), C₀t filtration (CF), mRNA-derived cRNA, and AFLP methods were applied to three diverse maize inbred lines (B73, Mo17, and CML69) with three replications per line (36 Maize GeneChips). Our results indicate that these particular target preparation methods offer only modest power to detect SFPs with the Maize GeneChip. Most notably, CF and MF are comparable in power, detecting more than 10,000 SFPs at a 20% false discovery rate. Although reducing sample complexity to ~125 Mb by AFLP improves SFP scoring accuracy over other methods, only a minimal number of SFPs are still detected. Our findings of residual repetitive DNA in labeled targets and other experimental errors call for improved gene-enrichment methods and custom array designs to more accurately array genotype

² M. Gore, P. Bradbury, R. Hogers, M. Kirst, E. Verstege, J. van Oeveren, J. Peleman, E. Buckler, and M. van Eijk. 2007. Evaluation of Target Preparation Methods for Single Feature Polymorphism Detection in Large Complex Plant Genomes. *The Plant Genome* S2: S-135-S-148.

large, complex crop genomes.

INTRODUCTION

Modern cultivated maize (*Zea mays* L.) boasts more genetic diversity than any other domesticated grass, retaining on average more than two-thirds of the nucleotide diversity of its wild relatives (Gaut et al., 2000; Tenaillon et al., 2001; White and Doebley, 1999). Indeed, DNA sequences of any two maize inbred lines differ from one another at an estimated frequency of a single nucleotide polymorphism (SNP) per 70 bases (silent sites) (Tenaillon et al., 2001). Considering such high levels of nucleotide diversity and a genome roughly equivalent in magnitude to the human genome (Arumuganathan and Earle, 1991), this yields about 30 million segregating sites. Intragenic linkage disequilibrium (LD) rates decline to minimal levels within two Kilobases (Kb) for a genetically diverse sample of tropical and temperate maize inbred lines (Remington et al., 2001). Due to this rapid breakdown of LD in a highly variable genome, an estimated one million SNP markers are required for genome-wide association studies.

Although the maize genome is a sizable ~2.5 Gigabase (Gb), the vast majority consists of several classes of retroelements known as long-terminal repeat (LTR)-retrotransposons (SanMiguel et al., 1996). LTR-retrotransposons are generally recombinationally inert, thereby confining most meiotic recombination to the gene rich or low-copy-number regions of the maize genome (Fu et al., 2002; Fu et al., 2001; Yao et al., 2002). Association mapping approaches, which rely on historical recombination for resolving complex traits, require that these regions of active recombination be identified and tagged. Because gene expression microarrays consist of oligonucleotides (oligos) designed from the sequence of expressed genes, they offer one potentially powerful means of genotyping thousands of recombinationally active

gene regions in parallel. The genotyping of sequence polymorphisms with an expression array is based on the concept that a perfectly matched target binds to an oligo probe or feature with greater affinity than a mismatched target (Borevitz et al., 2003; Singer et al., 2006). If an individual oligo feature on an expression array shows a significant and reproducible difference in hybridization intensity between genotypes or strains, it can serve as a polymorphic marker or single feature polymorphism (SFP). The goal of this study was to test the feasibility of expression arrays for use in SFP detection in maize.

The efficacy of Affymetrix expression arrays for permitting highly accurate scoring of SFPs has already been demonstrated in relatively small genomes such as ~4 Megabase (Mb) bacteria (*Mycobacterium tuberculosis*) (Tsolaki et al., 2004), ~12 Mb yeast (*Saccharomyces cerevisiae*) (Winzeler et al., 1998), and ~135 Mb *Arabidopsis thaliana* (hereafter *Arabidopsis*) (Borevitz et al., 2003). Expression arrays hybridized with DNA have also been used to map genetic loci and dissect traits (Singer et al., 2006; Steinmetz et al., 2002; Werner et al., 2005; Wolyn et al., 2004). Such whole-genome hybridization, however, has had limited success for detection of SFPs in crop plants with larger, more complex genomes, such as ~5.2 Gb barley (*Hordeum vulgare* L.) (Rostoks et al., 2005) and ~2.5 Gb maize (Kirst and Buckler, unpublished data, 2004). Thus, a target preparation method based on gene enrichment or complexity reduction is needed to exploit this potentially powerful technology.

One reasonably effective strategy is to score SFPs with cRNA derived from the less complex mRNA fraction of barley and maize (Cui et al., 2005; Kirst et al., 2006; Rostoks et al., 2005). Using cRNA as a surrogate for genomic DNA, however, has several notable limitations, including a requirement for extensive replication (e.g., 6X in Kirst et al., 2006) and a need to sample multiple tissues due to spatial and temporal expression of genes (e.g., 3X of six tissue types in Rostoks et al., 2005).

Methylation filtration (MF) with the bacterial *McrBC* restriction-modification system and C_0t filtration (CF) are two gene-enrichment technologies that have enabled a significant proportion of the maize gene space to be sequenced (Palmer et al., 2003; Whitelaw et al., 2003; Yuan et al., 2003). CF and MF yielded a four- to seven-fold enrichment in maize gene sequences as compared to control libraries (Rabinowicz et al., 1999; Yuan et al., 2003). MF exploits the differential methylcytosine patterns between genes and retrotransposons in plants. Unlike mammalian retrotransposons, those in plants are more heavily methylated than the rest of the genome (Rabinowicz et al., 2003; Rabinowicz et al., 2005). When plant retrotransposon DNA containing methylcytosine on one or both strands is preceded by a purine (G/A) residue (Raleigh, 1992; Sutherland et al., 1992), it is cleaved by *McrBC*, a novel type I GTP-dependent restriction endonuclease. This results in gene rich regions being digested much less frequently than retrotransposon blocks—a characteristic that has been used to clone and sequence the unmethylated portion (gene space) of genomes from several plant genera (Bedell et al., 2005; Palmer et al., 2003; Rabinowicz et al., 1999; Rabinowicz et al., 2005).

The principle underlying CF is based on the renaturation kinetics of DNA (Britten and Kohne, 1968) and has been used to differentially fractionate plant genomes according to copy number and base composition (Geever et al., 1989; Hake and Walbot, 1980; Peterson et al., 2002a; Yuan et al., 2003). Mechanically sheared genomic DNA is denatured and reassociated to a calculated C_0t value, a product of nucleotide concentration and reassociation time (Peterson et al., 2002a). The unrenaturated genome fraction enriched for low-copy number and genic sequences (High- C_0t) is then cloned and sequenced, while the renaturated moderately (Medium- C_0t) and highly repetitive (Low- C_0t) DNA fractions are excluded (Peterson et al., 2002a; Yuan et al., 2003).

A final technique, AFLP, uses the random distribution of restriction endonuclease recognition sites across a genome to make amplification libraries (Vos et al., 1995). By carefully selecting enzyme motifs and varying the number of selective bases in the amplification primers, it is possible to modulate both the number of unique, amplified fragments as well as genome complexity. Although standard AFLP procedures are not biased to gene regions, different random pools of DNA can be preferentially amplified and genotyped on expression arrays by changing enzymes. AFLP offers the additional advantage of being reproducible and amenable to high throughput processing.

Due to large amounts of repetitive, mobile DNA, the maize genome requires a target preparation method that offers both a high level of gene-enrichment and accurate scoring of SFPs. The objectives of this paper are to: (i) determine which target preparation method (CF, MF, mRNA, or AFLP) optimally enriches for gene sequences complementary to probe sequences on the Affymetrix GeneChip Maize Genome Array, and (ii) estimate SFP detection power for each target method.

MATERIALS AND METHODS

Sample and Array Specifications

To evaluate the effectiveness of several target preparation methods for detecting single feature polymorphisms (SFPs) in large, complex plant genomes, we conducted an experiment to score SFPs in three diverse maize inbred lines. Iowa Stiff Stalk Synthetic line, B73; non-stiff stalk line, Mo17; and tropical lowland CIMMYT (International Center for Maize and Wheat Improvement) line, CML69; represent the three major subpopulation groups of maize inbred lines (Liu et al., 2003; Remington et al., 2001). The Affymetrix Gene Chip Maize Genome Array (“Maize GeneChip”) has 17,555 probesets with 263,026 probe pairs for expression profiling 14,850 maize

genes (13,339 unique). Of the 17,555 probesets, 17,477 have 15 probe pairs, while the remaining 78 probesets have 14 or less probe pairs. Each probe pair consists of a perfect match (PM) probe and mismatch (MM) probe. The PM probe has a 25 bp sequence that is identical to a specific target gene transcript, whereas the MM probe differs from the PM probe by a single nucleotide substitution at the central base position. Array probes are designed from the sequence of expressed maize genes available in NCBI's GenBank (up to September 29, 2004) and *Zea mays* UniGene Build 42 (July 23, 2004) databases (<http://www.affymetrix.com>).

Target Synthesis and Array Hybridization

Total genomic DNA was extracted from powdered lyophilized leaf tissue using cetyltrimethylammonium bromide (CTAB) extraction buffer according to the protocol described by Saghai-Marooof et al. (1984). DNA was extracted in triplicate from a single genotyped tissue source, thus all DNAs isolated from the same inbred tissue source are technical replicates.

The maize genome was methylation filtered (MF) using McrBC as previously described by Zhou et al. (2002), with minor modifications. McrBC fragments were generated by incubating 60 µg genomic DNA with 600 U of McrBC (New England Biolabs, Ipswich, MA) at 37°C for 8 h, followed by heat inactivation of the enzyme at 65°C for 20 min. McrBC fragments ranging in size from ~12 Kb to less than 100 bp (data not shown) were separated on a low-melting 0.8% SeaPlaque® Agarose gel (Cambrex Bio Science Rockland, Inc., Rockland, ME). Most unwanted, restricted methylated DNA migrated to positions below the 1 Kb marker. Fragments \geq 1 Kb were excised from the gel and purified using the QIAEX II Gel Extraction Kit (Qiagen, Valencia, CA), according to the manufacturer's protocols.

C_0t filtration (CF) involved selecting the High- C_0t (HC) single-stranded (ss) DNA fraction as described by Peterson et al. (2002a). In brief, 50 μ g of genomic DNA was sheared to an average fragment size of 450 bp using a Misonix Sonicator 3000 (Misonix, Inc., Farmingdale, NY) with full power settings, for 24 cycles of 30 s of sonication and 1 min of cooling. Cations were removed using a Chelex ion-exchange column, followed by concentration and resuspension of the DNA in 0.5 M sodium phosphate buffer (SPB). DNA was transferred to capillary tubes, denatured in boiling water for 10 minutes, and allowed to renature to a C_0t value of 262 M•sec. A C_0t value is the product of the sample's nucleotide concentration (moles of nucleotides per liter), its reassociation time in seconds, and a buffer factor based upon cation concentration (Peterson et al. 2002a). Renatured DNA was then transferred to a hydroxyapatite (HAP) column (Bernardi, 1971) equilibrated with 0.03 M SPB. Finally, HC ssDNA was eluted by loading the HAP column with 0.12 M SPB.

Amplification of AFLP fragments was carried out according to the protocol described by Vos et al. (1995), using 200 ng genomic DNA as starting material. Sequences of the *TaqI* adapter were 5'-CTCGTAGACTGCGTAC-3' and 5'-CGGTACGCAGTCT-3', and sequences of the *MseI* adapter were 5'-GACGATGAGTCCTGAG-3' and 5'-TACTCAGGACTCA-3'. Sequences of the *TaqI*+A, *MseI*+C and *MseI*+G primers were 5'-GTAGACTGCGTACCGAA-3', 5'-GATGAGTCCTGAGTAAC-3' and 5'-GATGAGTCCTGAGTAAG-3', respectively. AFLP products were purified by standard sodium acetate/ethanol precipitation and dissolved in $T_{10}E_{0.1}$.

A total of 300 ng purified HC ssDNA, MF DNA or purified AFLP product were biotin-labeled in triplicate using the BioPrime[®] DNA labeling system (Invitrogen, Carlsbad, CA), as described by Borevitz et al. (2003). Specifically, 60 μ l 2.5X random octamer primers and 300 ng DNA were denatured in a total volume of

132 μ l at 95°C for 10 minutes and cooled on ice to allow annealing of random primers. Next, 15 μ l 10X dNTP/biotin-14-dCTP and 3 μ l Klenow fragments were added for primer extension and incubated overnight at 25°C. Labeled fragments were purified by standard sodium acetate/ethanol precipitation and dissolved in 30 μ l T₁₀E_{0.1}. For the labeled AFLP samples, a total of 15 μ g *Taq*I+1(A)/*Mse*I+1(C) and 15 μ g *Taq*I+1(A)/*Mse*I+1(G) from each sample were pooled and enough T₁₀E_{0.1} was added to bring the final volume to 30 μ l. The combination of these two AFLP +1/+1 samples was intended to represent an approximately 125 Mb fraction of the maize genome, which is almost equal in size to the *Arabidopsis* genome. These primer-enzyme combinations, however, are not optimized to specifically target gene regions.

Total RNA from homogenized frozen 4-week old leaf tissue was isolated using TRIZOL[®] reagent (Invitrogen, Carlsbad, CA) and Qiagen RNeasy Columns (Qiagen, Valencia, CA) according to the manufacturers' protocols. Total RNA was isolated from harvested leaves of individual plants, thus all RNAs isolated from a specific inbred are biological replicates. A total of 7 μ g of each RNA sample was used for double-stranded cDNA synthesis and biotin-labeling of antisense cRNA, as described in the manual accompanying GeneChip Expression 3'-Amplification Reagents One-Cycle cDNA Synthesis Kit and One-Cycle Target Labeling Assay (Affymetrix, Santa Clara, CA). Finally, 15 μ g biotin-labeled cRNA per reaction was supplemented with T₁₀E_{0.1} to achieve a final volume of 30 μ l.

Hybridizations on GeneChip Maize Genome Arrays (Affymetrix, Santa Clara, CA) were carried out by an Affymetrix service station (ServiceXS, Leiden, The Netherlands), according to Affymetrix protocols. In total, 36 GeneChips were used in this study. Three technical replicates of CF, MF, and AFLP for each line were hybridized to 27 GeneChips, and three biological replicates of mRNA for each line were hybridized to 9 GeneChips.

GeneChip Quality Control

The scanned image of each GeneChip was visually inspected for spatial artifacts using the method Image of the *affy* package (<http://www.bioconductor.org>) in the freely available statistical package R (<http://www.r-project.org>; Ihaka and Gentleman, 1996). Standard Affymetrix quality control parameters for assessing arrays were checked and determined to be reasonably concordant with the manufacturer's recommendations (Gene-Chip Expression Analysis Data Analysis Fundamentals; <http://www.affymetrix.com>).

Pearson's correlations of raw PM probe intensities between arrays of the same target preparation method ranged from 0.95 to 0.99 within line, while between lines correlations were in the range of 0.85 to 0.95. Notably, our analysis revealed that one of the Mo17 line-CF replicates had low correlations (0.5 to 0.6) to the other CF lines and replicates. Therefore, we excluded this outlier array from all further analyses. The inbred line assignment for each GeneChip was further verified by analyzing the average Euclidean distance between standardized \log_2 probe intensities of 289 probesets. All quality control statistical analyses were carried out using SAS (SAS Institute, Cary, NC, USA). The PROC CORR and PROC DISTANCE statements were used to calculate correlations and distances, respectively.

Maize Sequence Validation Datasets

Methodology

A dataset for validation of detected SFPs was created from sequence alignments that matched the sequence of probes on the Maize GeneChip (Maize_probe_tab.txt; <http://www.affymetrix.com>). Specifically, the 25 bp nucleotide sequence of each PM probe was compared to a 25 bp sliding window of nucleotide sequence along all B73, Mo17, and CML69 sequence alignments in the Panzea

database (<http://www.panzea.org>) (Zhao et al., 2006). The reverse complement of each PM probe sequence was also used to search Panzea. If an exact match between an alignment and PM probe sequence was identified for at least one of the lines, a 25 bp string initiated from the probe start position within the alignment was extracted for all three lines. All three extracted 25 bp strings were then aligned to the initial queried PM probe sequence. This allowed for the number of exact match nucleotides to be counted and the position of any SNPs within the string to be recorded. Any extracted string containing a gap (insertion or deletion) or ambiguous nucleotide was discarded. The resulting sequence dataset contained all B73, Mo17, and CML69 sequences from Panzea that exactly matched Affymetrix PM probes for at least one of the inbred lines, along with any corresponding mismatch sequences from the remaining lines.

Additional criteria were used to help ensure the quality of sequences in the SFP validation dataset. For example, many of the alignments included two sequencings of B73 and Mo17 for quality control. If the two B73 strings or the two Mo17 strings were not identical for any 25 bp nucleotide sequence, the sequence at that position was not used. Also, on rare occasion (<0.5%) one of the lines was found to have more than four SNPs when compared to the probe sequence. Sequence at that location was excluded from the dataset, as these SNPs may have been caused by an alignment error rather than actual sequence variation.

Primary SFP Validation Dataset

The primary SFP validation dataset was used to calculate SFP detection power for each target preparation method. This validation dataset contains 38,259 sequences of 25 bp (~1 Mb) from B73, Mo17, and CML69 for 14,651 PM probes, of which 1,620 probes (11%) detect one to four SNPs in at least one of the three maize inbred lines. There are a total of 1,998 segregating sites (S), which translates to a θ_{PMprobe}

estimate of 0.0014. The number of SNPs detected by a PM probe in each inbred line is as follows: B73, 453; Mo17, 1070; and CML69, 802. Of the 14,651 PM probes with available sequence data for a maize inbred line, there are a maximum of 32,511 pairwise probe comparisons and 2,677 (8.2%) of these involve a PM probe that detects at least one SNP—potentially leading to the detection of 2,677 SFPs. The calculated SFP rate in this dataset for each inbred pairwise probe comparison is as follows: B73-CML69, 7.9% (742/9,386); B73-Mo17, 8.3% (1,128/13,631); and CML69-Mo17, 8.5% (807/9,494). Consequently, with this dataset we can detect at most 2,677 SFPs with each target preparation method if all 14,651 PM probes are members of probesets called Present (detected) by the Affymetrix Microarray Suite version 5 (MAS5) algorithm (Liu et al., 2002) on all CF, MF, mRNA, or AFLP arrays.

The observed SNP diversity ($\theta_{\text{PMprobe}}=0.0014$) in the primary SFP validation dataset is about 19% of the SNP diversity ($\theta_{\text{PMprobe}}=0.0075$) reported by Kirst et al. (2006) when PM probes were used to genotype a diverse set of maize inbred lines. In Kirst et al. (2006), cRNA was hybridized to an 8K Maize CornChip0, which contains probes that were designed from the sequence of a limited number of maize genotypes (e.g., ~50% B73 sequence). Unlike the Maize CornChip0, probes on the Maize GeneChip were designed to be robust for multiple maize genotypes by masking polymorphisms identified in the expressed sequences of over 100 maize lines (<http://www.affymetrix.com>; Stupar and Springer, 2006). Therefore, probes on the Maize GeneChip were systematically designed to hybridize regions of gene transcripts with lower than average levels of nucleotide diversity, and as such, resulted in low rates of SNP detection in this study.

Secondary SFP Validation Dataset

The secondary SFP validation dataset was used to calculate SFP detection

power in an unbiased manner. This secondary dataset, a subset of the primary SFP validation dataset, was constructed with only PM probes from probesets that were called Present by MAS5 on all CF, MF, and mRNA arrays. AFLP was not analyzed with the secondary SFP validation dataset due to the low number of shared probesets called Present by MAS5 on AFLP arrays. The secondary SFP validation dataset contains 23,873 sequences of 25 bp (~0.6 Mb) from B73, Mo17, and CML69 for 9,039 PM probes, of which 835 PM probes (9.2%) detect one to four SNPs in at least one of the three maize inbred lines. With the 9,039 PM probes there are 20,666 pairwise probe comparisons and 1,409 (6.8%) of these could potentially detect an SFP.

Polymorphic Probeset Validation Dataset

We also investigated whether probesets (probeset level analysis) containing one or more polymorphic probes (polymorphic probesets) are detected with greater accuracy than SFPs (probe level analysis). A dataset for validation of detected polymorphic probesets was constructed using probesets for which all probe sequences and SNPs were known. In the SFP validation dataset described above, very few probesets had all 15 probes match a sequence in the Panzea database. To construct a dataset of probesets with no missing sequence data, we first identified probesets that were called Present by the MAS5 algorithm on all CF, MF, and mRNA arrays. Second, probesets with eight or more probes matching an alignment sequence were identified. Third, probes within those probesets that had no matching Panzea sequence were removed from the dataset. The resulting probeset validation dataset contained 289 probesets, each consisting of between eight to fifteen probes. Of these 289 probesets, a total of 109 (38%) contained at least one mismatch probe due to a SNP in one of the three lines and as such were defined as polymorphic.

Hybridization Data Pre-Processing and Normalization

Raw CEL files were background corrected (RMA; Irizarry et al., 2003) and then normalized (Quantiles; Bolstad et al., 2003). We found that processing the hybridization data with RMA and Quantiles resulted in equivalent or higher SFP detection power as that obtained with the spatial correction method described in Borevitz et al. (2003). MAS5 was used to remove probesets called Absent or Marginal (unreliably detected) before probe level analysis. Probesets were retained for further analysis if called Present (detected) for a method specific set of nine GeneChips (MF, mRNA, and AFLP) or eight GeneChips (CF). RMA, Quantiles, and MAS5 methods of the *affy* package were carried out in R.

Detecting SFPs in Hybridization Data

SFPs were identified in pre-processed hybridization data using the two-step strategy mixed model as described in detail by Kirst et al. (2006). Analyzed datasets of background, normalized probe intensities were derived from probesets called Present that included at least one probe sequence in common with the SFP validation dataset. Each probeset was analyzed separately. The overall array mean for each array was subtracted from the \log_2 of the probe intensity. The following mixed model was fit to the resulting values in SAS:

$$I_{ijk} = L_i + P_j + a_{ik}(L_i) + L_i * P_j + e_{ijk},$$

where $I_{ijk} = \log_2(\text{probe intensity})$ for the i th maize inbred line for the j th probe on the k th array of that line less the array mean for the probeset,

L_i = the effect of the i th line,

P_j = the effect of the j th probe,

$a_{ik}(L_i)$ = the effect of the k th array of the i th line nested within line,

$L_i * P_j$ = the interaction effect for the i th line and the j th probe; represents

SFPs,

e_{ijk} = random error,

$i = 1, 2, 3$, the number of lines,

$j = 1, \dots, 15$, the number of probes in a probeset, and

$k = 1, 2, 3$, the number of arrays per line.

The data were analyzed using SAS PROC MIXED, fitting line and probe as fixed effects and array as a random effect nested in line using the following model statements:

model intensity = line probe line*probe;

random array / subject = line;

lsmeans line*probe / diff;

The LSMEANS statement in SAS was used to generate pairwise comparisons between inbred lines at each probe with a t-test of the null hypothesis that the difference was zero. A statistically significant non-zero value indicated a potential SFP. All pairwise t-test comparisons were performed in one of two ways: using the standard error from the probeset as indicated in the model above (probeset error term t-test) or assuming a constant error term from the complete array (array error term t-test).

The SFP validation sets were used to confirm whether detected SFPs were true or false positives, thereby allowing for the estimation of detection power at empirically calculated false discovery rates (FDRs). To do this, comparisons between lines at each probe were first sorted by p-value. For each p-value, the FDR was calculated as the number of comparisons with an equal or lower p-value that were false SFPs divided by the total number of comparisons with an equal or lower p-value. The power was calculated as the number of true SFPs with an equal or lower p-value divided by the total number of true SFPs in the dataset. Calculations were performed for both the probeset error term t-test as well as the array error term t-test.

All R and SAS scripts, raw GeneChip data, sequences for validation set probes, and lists of identified SFPs are available upon request. Raw GeneChip data will also be deposited in PLEXdb (Plant Expression Database; <http://plexdb.org/>).

RESULTS

Probe Performance

More than 14,000 perfect match (PM) probe sequences on the Affymetrix GeneChip Maize Genome Array (“Maize GeneChip”) are an exact match to at least one of the three maize inbred line sequences in the Panzea database. In these cases, we expect PM probe signal intensity to be greater than that of the mismatch (MM) probe. However, there are many instances where the MM probe has higher signal intensity than the PM probe (MM>PM) in spite of the fact that the PM probe is known to be a perfect match for the target. Figure 2.1 shows the distribution of PM probe signal intensities and indicates the portion for which the MM signal exceeds the PM signal. PM signal on AFLP and mRNA GeneChips is strongly skewed toward the lower end of the \log_2 PM intensity range, while CF and MF exhibit a more normal distribution. As the signal intensity of PM probes on mRNA GeneChips increases, the proportion of MM>PM probe pairs drastically diminishes. In comparison, the proportion of probe pairs on MF and CF GeneChips where the MM probe signal intensity is greater than the PM probe is more uniform across the \log_2 PM signal intensity range.

Analysis of PM and MM probe signal intensity data from mRNA GeneChips confirms that when gene expression levels are high, signal from target sequence overwhelms the noise from non-target sequences cross-hybridizing to probes. The main problem with mRNA samples is that they contain transcripts from genes with low levels of expression (i.e., signal near background levels) that in many cases makes SFP detection difficult. In contrast, CF and MF samples most likely contain low to

intermediate levels of repetitive DNAs that are spuriously annealing to probes across all PM intensity levels. On the other hand, the most important factor affecting AFLP samples is that they are not well represented by GeneChip probes.

Array Coverage of Gene Enrichment Methods

Because of the significant number of identified MM>PM probe pairs, we used the Affymetrix Microarray Suite version 5 (MAS5) algorithm to filter hybridization data so that data for probesets unreliably detected could be eliminated. The MAS5 algorithm uses probe pair data in a Wilcoxon signed rank test to determine whether PM probes have a higher hybridization intensity signal than their analogous MM probes (Liu et al., 2002). Depending on the outcome of this test, one of three detection calls (Present, Absent, or Marginal) is assigned to each probeset. We performed a separate MAS5 analysis on each GeneChip. Hybridization data were maintained if probesets were called Present for each GeneChip in a target preparation method set, while data from probesets called Marginal or Absent were removed from further analyses.

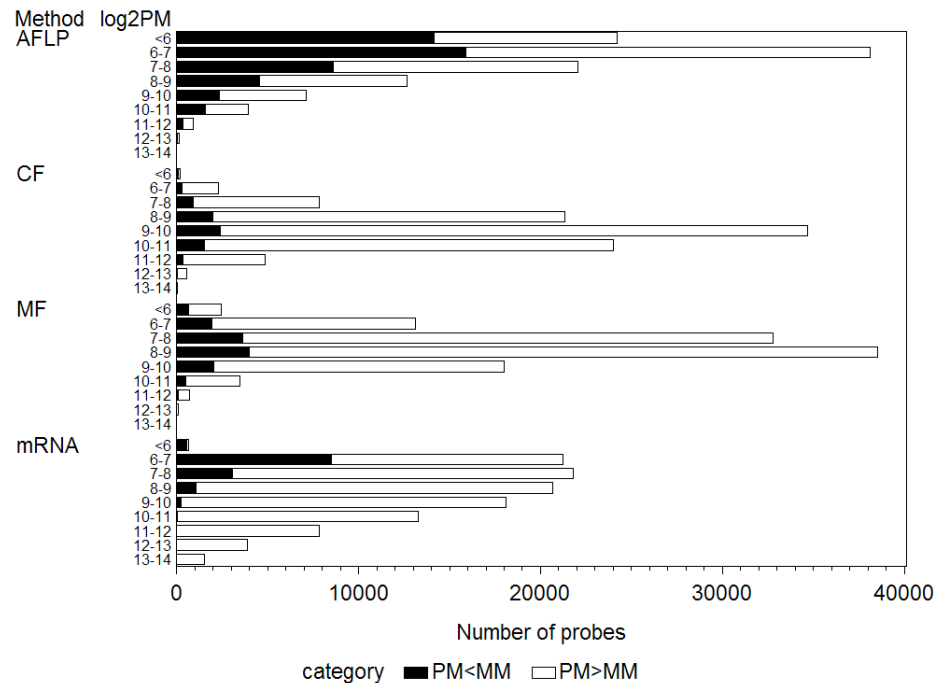


Figure 2.1. Frequency distribution of perfect match (PM) and mismatch (MM) probe pair signal intensity ratios. Probe pair signal intensity ratios are shown according to log₂PM range for AFLP, C₀t filtration (CF), methyl filtration (MF), and mRNA.

Although the primary purpose for employing the MAS5 algorithm was to increase the ratio of true positive to false positive SFPs (i.e., decrease Type I error rate), this analysis also allowed us to calculate the total number of probesets called Present for GeneChips of each target preparation method. Because probes are designed from the sequence of expressed maize genes, the number of probesets called Present serves as a direct indicator of how well each method provides sequences complementary to probes on the Maize GeneChip. The number of probesets called Present by MAS5 differs substantially by target preparation procedure: AFLP, 646 (4%); mRNA, 9,661 (55%); MF, 12,975 (74%); and CF, 14,895 (85%). CF and MF provide for a greater representation of complementary gene sequences than mRNA

fractions isolated from a single tissue type (leaf) and specific developmental stage (V4-5). A larger portion of the maize gene space is sampled by CF and MF, while transcript presence and location are dependent on the temporal and spatial pattern of gene expression. AFLP has more than 10-fold fewer Present calls, suggesting that the selected restriction enzymes (*TaqI* and *MseI*) and amplification protocol substantially reduce maize genome complexity without highly enriching for gene fragments complementary to array probes.

Assessment of Power to Detect SFPs

In order to estimate SFP detection power afforded by CF, MF, mRNA, and AFLP, we first constructed a primary SFP validation dataset containing all B73, Mo17, and CML69 sequences from the Panzea database that matched to a PM probe sequence (see detailed description in Materials and Methods under Maize Sequence Validation Dataset). We determined that 1,620 out of the 14,651 validation dataset probes should detect one to four SNPs (SNP probes) in at least one inbred line. The other 13,031 probes in the SFP validation dataset should not detect any SNPs when hybridized to target sequences from any of the three inbred lines (non-SNP probes). Of the possible 32,511 pairwise probe comparisons between B73, Mo17, and CML69, there are 2,677 comparisons that could potentially detect an SFP. The number of SNP and non-SNP validation dataset probes contained within probesets called Present by MAS5 was determined for each target preparation method (Table 2.1). The number of detected SNP and non-SNP probes shared with the primary SFP validation dataset is highest for CF and MF, which is reflective of their overall success in enriching for genes represented as probes on the array. Subsequently, we calculated the total number of potential SFPs that could be identified through pairwise probe comparisons of all three lines with MAS5 detected SNP and non-SNP probes (Table 2.1). CF and

MF provide for a greater representation of probes on the GeneChip and in the SFP validation dataset, and as such, have the potential to provide more opportunities to detect SFPs.

Table 2.1. SFP detection potential of target preparation methods.

Method†	Non-SNP Probes‡		SNP Probes§		SFPs¶	
	No. Probes#	%††	No. Probes	%	No. SFPs‡‡	%§§
CF	12197	94	1430	88	2429	91
MF	10851	83	1202	74	2009	75
mRNA	9285	71	1019	63	1707	64
AFLP	229	2	26	2	38	1
Total	13031	100	1620	100	2,677	100

† CF, *Cot* filtration; and MF, Methyl Filtration.

‡ non-SNP Probes, primary validation dataset probes that should not detect a single nucleotide polymorphism (SNP) in B73, Mo17, and CML69.

§ SNP Probes, primary validation dataset probes that should detect anywhere from 1 to 4 SNPs in B73, Mo17, and/or CML69 (not all three).

¶ SFPs, individual probes that should detect at least 1 SNP in B73, Mo17, and/or CML69 (not all three), and therefore, have potential for being detected as polymorphic markers or single feature polymorphisms (SFPs) in pairwise probe comparisons.

Number of validation non-SNP and SNP probes that are contained within a probeset called Present by MAS5.

†† Percentage of all non-SNP or SNP probes in the primary validation dataset that are members of probesets called Present by MAS5 on arrays of each target preparation method.

‡‡ Number of pairwise probe comparisons that could potentially detect an SFP. These calculated numbers are based on the specific MAS5 detected non-SNP and SNP probes for each target preparation method and are cumulative across the three inbred lines (i.e., B73 vs. Mo17; B73 vs. CML69; or Mo17 vs. CML69). Not all inbred line pairwise comparisons were possible for some probes, because sequence information was missing for one of the lines.

§§ Percentage of all SFPs represented in the primary validation dataset that are detectable on arrays of each target preparation method.

We applied a mixed model to background, normalized probe intensity data from all probes of probesets called Present that share at least one probe sequence in common with the SFP validation dataset. The mixed model accounts for line, probe, and probe-by-line effects, which are sources of variation in probe intensities and probeset signal estimates (Kirst et al., 2006). A significant negative interaction between a probe and one or more inbred lines suggests that at least one DNA sequence polymorphism is reducing the signal intensity of the probe. Significant probe-by-line

effects were detected using pairwise comparisons of individual probe intensity estimates between inbred lines, and SFP detection power was calculated at 5, 10, 20, 30, and 40% false discovery rates (FDRs) for each target preparation method (Table 2.2).

Table 2.2. Mixed model analysis of SFP detection power.

FDR†	CF‡		MF‡		mRNA§		AFLP‡	
	Power (%)	No. SFP#	Power (%)	No. SFP	Power (%)	No. SFP	Power (%)	No. SFP
5%††	1	26	1	20	-	-	45‡‡	17
10%	3	78	14	284	2	29	45	17
20%	30	734	26	514	26	447	45	17
30%	41	1002	37	736	34	573	45	17
40%	49	1179	43	869	39	662	45	17

† FDR, false discovery rate.

‡ Array error term t-test.

§ Probeset error term t-test.

¶ SFP detection power was calculated as the proportion of detected true positive SFPs to the total number of potential SFPs at an empirically determined FDR using the primary SFP validation dataset. Details as to how empirical FDRs were determined are provided in the Materials and Methods under Detecting SFPs in Hybridization Data.

Number of SFP detected at an empirically determined FDR.

†† Power estimates at 5% FDR are less statistically reliable due to the lower number of detected SFPs.

‡‡ All power estimates for AFLP were based on a low number of observations and are therefore less statistically reliable than those for the other methods.

Power to detect SFPs was calculated as the proportion of detected true positive SFPs (sequence confirmed) to the total number of expected SFPs (Table 2.1) at an empirically determined FDR. SFP detection power for mRNA was calculated using the probeset error term and for the other three methods was calculated using the array error term. SFP detection power at 5% FDR is almost negligible (1%) for both MF and CF. MF does detect 284 confirmed SFPs at 10% FDR, while the number of confirmed SFPs detected by CF at higher FDRs exceeds all other methods. SFP detection power of mRNA and MF are almost equivalent at FDRs of 20% and higher, but more SFPs are detected using MF by virtue of its greater probe coverage. AFLP scores SFPs with

more accuracy than the other target preparation methods, but the numbers of detected SFPs are far lower due to AFLP's inferior SFP detection potential with this particular GeneChip design. This low potential directly results from the primary amplification of non-genic random sequences. Interestingly, no additional SFP detection power is gained until 60% FDR with AFLP, as power is static at 45% from 5 to 40% FDRs. A likely explanation is that AFLP accurately scores all the SFPs for the few genes that it can at 5% FDR with limited cross-hybridization from other amplified targets.

The mixed model was also applied to a subset of the probe intensity data that consists of 1,440 probesets called Present on all CF, MF, and mRNA GeneChips. All of the parsed probesets have one or more probe sequences in common with the secondary SFP validation dataset (see detailed description in Materials and Methods under Maize Sequence Validation Dataset). The secondary validation dataset of shared probes contains 8,204 non-SNP probes and 835 SNP probes (9,039 total probes). Of the 20,666 possible pairwise probe comparisons, there is potential to detect 1,409 SFPs. Analysis of the shared probes dataset enabled us to compare the SFP detection power of each method without any probeset biases, because all of the analyzed validation probesets had signal intensities greater than background on all CF, MF, and mRNA GeneChips. AFLP was not included in the shared probes analysis due to the low number of validation probes shared with the other three methods. The results of the shared probes analysis (Table 2.3) are similar to those of the initial complete datasets (Table 2.2), with the exception that a reduction in probe numbers eliminated SFP detection power at 5% FDR for CF. In addition, based on results presented in Tables 2.2 and 2.3, SFP detection power is reduced 10% at 10% FDR for MF in the shared probeset analysis. These observed losses of power are mainly due to the removal of probes from the complete validation dataset that detected true positive SFPs (5 to 10% FDR) on CF and/or MF GeneChips.

Table 2.3. Mixed model analysis of SFP detection power with shared probes.

FDR	CF†		MF†		mRNA‡	
	Power (%)§	No. SFP	Power (%)	No. SFP	Power (%)	No. SFP
5%¶	-	-	1	21	-	-
10%¶	3	36	4	53	1	9
20%	34	475	23	330	24	337
30%	42	598	35	498	33	462
40%	48	680	43	603	38	536

† Array error term t-test.

‡ Probeset error term t-test.

§ SFP detection power was calculated as the proportion of detected true positive SFPs to the total number of potential SFPs at an empirically determined FDR using the secondary SFP validation dataset. Details as to how empirical FDRs were determined are provided in the Materials and Methods under Detecting SFPs in Hybridization Data.

¶ Power estimates at 5 and 10% FDRs are less statistically reliable due to the lower number of detected SFPs.

SNP Position Effect

Results of SFP detection power reported in Table 2.2 indicate that with any one of the target preparation methods a large proportion (51 to 61%) of potential SFPs resulting from SNPs remain undetected. The location of a SNP within the 25 bp probe affects target binding efficiency and in so doing also affects PM probe signal intensity. SNP position is defined as the position from the edge of the probe. Position 1 is the first base at either end, and position 13 is the center of the probe. SNPs within the internal 15 bases (positions 6 to 20) have been found to reduce hybridization much more than nucleotide mismatches within the external 5 bases (positions 1 to 5 and 21 to 25) (Kirst et al., 2006; Ronald et al., 2005; Rostoks et al., 2005).

We investigated the impact of SNP position on SFP detection for 984 probes that recognize only a single SNP upon hybridizing to the B73, Mo17, and/or CML69 target sequence on CF, MF, and mRNA GeneChips. Of the 984 probes in the probeset dataset, 38% (376) and 62% (608) detect an edge SNP and internal SNP, respectively. The percentage of detected and undetected SFPs resulting from either edge or internal SNPs was calculated (Table 2.4). Detected SFPs (78 to 85%) are primarily the result

of internal SNPs, whereas undetected SFPs represent an approximate 1:1 ratio of edge-to-internal SNPs. Thus, as expected, the data summarized in Table 2.4 show that SFPs are called more often if the SNP occurs in the internal region. Also, the percentage of detected SFPs resulting from an edge SNP increases as FDR approaches 40%. SNP position effects are similar for CF, MF, and mRNA. We also examined whether probes detecting multiple SNPs (2, 3, or 4 SNPs) are detected at the same rates as probes detecting a single SNP. Based on analyzed SFP data, the former are called as SFPs no more or less frequently than the latter (data not shown).

Table 2.4. The distribution of SNP position in probes that detect a single SNP.

FDR	Position†	CF		MF		mRNA	
		No. Probe	%	No. Probe	%	No. Probe	%
5%	Edge	0	0	1	8	0	0
	Internal	0	0	12	92	0	0
10%	Edge	0	0	4	17	0	0
	Internal	16	100	19	83	2	100
20%	Edge	61	19	30	16	23	10
	Internal	258	81	152	84	203	90
30%	Edge	22	24	32	24	16	18
	Internal	69	76	101	76	72	82
40%	Edge	19	32	28	37	17	27
	Internal	41	68	47	63	46	73
Total‡ Detected	Edge	102	21	95	22	56	15
	Internal	384	79	331	78	323	85
Total Not Detected	Edge	274	55	281	50	320	53
	Internal	224	45	277	50	285	47

† Edge: 1 to 5 bp or 21 to 25 bp SNP position within probe. Internal: 6 to 20 bp SNP position within probe.

‡ Probe position distribution for combined total detected and not detected dataset: Edge 38% (376) and Internal 62% (608).

Detection Rate of Polymorphic Probesets

Alternatively, we examined whether it is more effective to identify probesets (probeset level analysis) that contain one or more polymorphic probes rather than

individual SFPs (probe level analysis). One rationale for this analysis is that as the number of polymorphic probes in a probeset increases, so does the difficulty of identifying specific SFPs. This difficulty stems from the fact that a target binds weakly when not identical to the probe. As a result, polymorphic probes do not provide an unbiased estimate of DNA (CF, MF, and AFLP) or gene expression levels (mRNA). And yet an accurate estimate of DNA or gene expression levels is required to determine which probes are polymorphic. In addition, a single probe comparison between two lines involves six data points, whereas a probeset comparison involves 90 data points. For these reasons, we hypothesized that a probeset analysis would be far more powerful than the analysis of individual probes.

To estimate the power to detect polymorphic probesets for CF, MF, and mRNA, we constructed a validation set of 289 probesets containing 8 to 15 probes with matching Panzea sequence, of which 109 (38%) contained at least one polymorphic probe (see detailed description in Materials and Methods under Maize Sequence Validation Dataset). AFLP was not included in the probeset level analysis due to the low number of AFLP probesets called Present and shared in common with the other three method's arrays. The intensity data for probes within these probesets were analyzed using the mixed model. The p-value from the F-test of probe by line interaction was recorded for each probeset and used to rank them in ascending order. Power to detect polymorphic probesets for the three target methods was calculated and summarized in Table 2.5. Irrespective of target preparation method, in this study Maize GeneChips are more effective in identifying polymorphic probesets than they are in detecting SFPs (Table 2.5). Compared to mRNA (19 to 68%), gain in power over SFP detection with CF (35 to 38%) and MF (22 to 43%) is not as dramatic, because DNA-based preparation methods should result in more normalized target copy number ratios. Even though the impact of poor DNA or gene expression level

estimates is minimized when detecting polymorphic probesets, one significant downside is that individual polymorphic probes are not identified as markers.

Table 2.5. Mixed model analysis of polymorphic probeset detection power.

FDR	CF			MF			mRNA		
	No. PP [†]	Power (%) [‡]	Power Gain (%) [§]	No. PP	Power (%)	Power Gain (%)	No. PP	Power (%)	Power Gain (%)
5%¶	-	-	-	-	-	-	21	19	19
10%	40	37	34	39	36	22	76	70	68
20%	72	66	36	56	51	25	82	75	49
30%	84	77	35	72	66	29	88	81	47
40%	95	87	38	94	86	43	90	83	44

[†] No. PP, Number of detected polymorphic probesets at an empirically determined FDR.

[‡] Polymorphic probeset detection power was calculated as the proportion of detected true positive SFPs to the total number of potential SFPs at an empirically determined FDR. Details as to how empirical FDRs were determined are provided in the Materials and Methods under Detecting SFPs in Hybridization Data.

[§] Power Gain, The percent gain in detection power was calculated as polymorphic probeset detection power (%) minus SFP detection power (%) in Table 2 at an empirically determined FDR.

[¶] Power estimates at 5% FDR are less statistically reliable due to the lower number of detected polymorphic probesets.

DISCUSSION

Conventional methods for SNP discovery in large-scale association mapping studies rely on resequencing candidate gene alleles across distinct individuals of a test population, followed by scoring known SNPs on individuals using one of several array-based SNP genotyping technologies (reviewed in Syvänen, 2005). Expression arrays, however, may offer a more rapid and cost-effective approach. Affymetrix GeneChip expression arrays hybridized with total genomic DNA have successfully functioned as both a polymorphism discovery and genotyping system in *Arabidopsis* and yeast (Hazen and Kay, 2003). Here, we tested whether the Affymetrix GeneChip is appropriate for highly parallel genotyping of larger, more complex genomes such as maize. The Maize GeneChip was evaluated as a high-density platform to detect SFPs

in cRNA or DNA hybridization data from three diverse maize inbred lines (B73, Mo17, and CML69).

Targets enriched for gene-content and/or reduced in genome complexity were generated by MF, CF, mRNA, and AFLP as a means to score SFPs across the retrotransposon-rich maize genome, but only modest SFP detection power was achieved when these targets were hybridized to the Maize GeneChip. For example, only 39% of expected SFPs were scored with cRNA at 40% FDR--far fewer than the previously reported ~70-80% of known sequence polymorphisms scored as SFPs using maize or barley cRNA (Cui et al., 2005; Kirst et al., 2006; Rostoks et al., 2005). The extent of GeneChip replication (Kirst et al., 2006; Rostoks et al., 2005), sampling of multiple tissues (Rostoks et al., 2005), and conservative five percentile cutoff (Cui et al., 2005) are the major experimental and data analysis demarcations leading to higher sensitivity in these other cRNA-based SFP studies. In the seminal *Arabidopsis* SFP work of Borevitz et al. (2003), at least 57% of known polymorphisms were detected at 13% FDR with labeled total genomic DNA as the target. Of the DNA-based methods evaluated here, MF, CF, and AFLP detected anywhere from 26 to 45% of SFPs at 20% FDR.

What factors are responsible for reducing SFP detection power in this study? Sequencing errors in the Panzea database may be one such factor, if such errors reduced overall detection power by generating undetectable false SFPs. Every effort, however, was made to filter out such sequencing errors before assessing power. As noted in previous SFP studies (Kirst et al., 2006; Ronald et al., 2005; Rostoks et al., 2005), we found that SFPs are detected more robustly if a nucleotide polymorphism in a target sequence binds within the internal 15 bases of the complementary PM probe, whereas edge SNPs are less frequently detected below 40% FDR. The actual minimization of power by this SNP position phenomenon was not quantified in this

study. The binding of spurious non-target repeat DNAs and multigene family member sequences to probes represents another potential source of genotyping error, compromising power and FDR. In addition, increasing the number of GeneChip replicates has been shown to improve power and FDR (Borevitz et al., 2003; Rostoks et al., 2005), and no doubt this study would have benefited from the same.

Despite the modest detection sensitivity when compared to SFP experiments using smaller genome species, this study marks the first report of using genome-filtered DNA targets to reliably identify more than 10,000 SFPs in a plant genome that contains at least 75% LTR-retrotransposons (San Miguel et al., 1996) and is 20X the size of *Arabidopsis*. Based on SNP diversity of maize sequences in the primary SFP validation dataset, we determined that 8.2% (2,677/32,511) of all pairwise probe comparisons involve a SNP probe (SFP diversity). Using the power results presented in Table 2.2 and measure of SFP diversity (0.082), we estimated the number of probes from probesets called Present (MAS5) that would be correctly identified as true SFPs on the Maize GeneChip (Table 2.6). We then analyzed probe intensity data from Present probesets with the mixed model to determine the observed number of SFPs detected on entire GeneChips. The p-value cutoffs from the primary SFP validation dataset were used to determine the number of detected SFPs at each FDR. The number of observed true SFPs was in turn calculated by multiplying the number of SFP detected by (1-FDR). The difference between the estimated and observed number of SFPs can be accounted for by the fact that the estimate of SFPs is founded on SNP diversity and does not include indel diversity, whereas observed SFP numbers account for indels. Kirst et al. (2006) reported that indels represent 40% of all polymorphisms occurring between PM probe and maize target gene sequences.

Table 2.6. Estimated and observed number of true SFPs that were identified on the whole Maize GeneChip.

	CF		MF		mRNA		AFLP	
	No. SFP		No. SFP		No. SFP		No. SFP	
FDR	Est.†	Obs.‡	Est.	Obs.	Est.	Obs.	Est.§	Obs.§
5%¶	549	1385	478	992	-	-	1072	918
10%	1647	3046	6698	10248	712	661	1072	1130
20%	16474	26646	12439	18454	9259	12702	1072	1056
30%	22515	35422	17701	25392	12108	15982	1072	1448
40%	26908	40729	20572	29726	13889	17054	1072	1493

† The estimated (Est.) number of true SFPs detected on the whole array was calculated by multiplying the probability (0.082) that a pairwise probe comparison involves a SNP probe, the power results shown in Table 2, and number of probes from probesets called Present by MAS5 for each target preparation method. The estimated number of true SFPs for the entire array is less than observed, because insertions/deletions (indels) and gene copy number differences are not taken into account.

‡ The observed (Obs.) number of true SFPs was determined for each target preparation method by analyzing probe intensity data from probesets called Present (MAS5) using the mixed model. The p-value cutoffs from the primary SFP validation dataset were used to determine the total number of detected SFPs on the entire array at each FDR. The number of observed true SFPs was in turn calculated by multiplying the number of SFP detected by (1-FDR).

§ Estimated number of SFPs for AFLP was based on a low number of observations and is therefore less statistically reliable than those for the other methods.

¶ The estimated number of detected SFPs at 5% FDR is less statistically reliable based on the SFP detection power results shown in Table 2.2.

In most cases, a 10% or lower FDR is acceptable when array genotyping individuals for an association study, but this is highly dependent on sample size, marker density, and levels of genome-wide LD. For example, MF is estimated to identify over 6,000 SFPs between the three maize inbred lines at 10% FDR, which results in a cost of ~\$0.38 per SFP (\$2250/9 arrays). After the initial investment to identify SFPs, the cost per SFP dramatically lowers to ~\$0.04 because subsequent genotyping requires only one array per individual (Borevitz et al., 2003). These estimated costs per SFP are very competitive to those reported for the ATH1 GeneChip (~\$0.30 per SFP and ~\$0.05 per SFP) in 2003 by Borevitz and colleagues. At 20% and higher FDRs, CF detects 1.3X more SFP than MF; however, these more liberal error rates are undesirable for most marker applications.

Although AFLP has far greater detection power from 5 to 20% FDRs, the

AFLP design tested here has inferior SFP detection potential and thus does not constitute an economical means of scoring SFPs on the Maize GeneChip. Even though the amplified target fraction contains about 5% of the maize genome (125 Mb/2500Mb), most amplicons are non-genic, random sequences which result in 4% of probesets called Present. On the other hand, CF and MF are highly preferable to labeling total genomic DNA for a large genome plant species (Rostoks et al., 2005; Buckler and Kirst, unpublished data, 2004) and are recommended for scoring SFPs when utilizing the Maize GeneChip. Compared to the other two methods, CF and MF not only provide for the highest coverage of array probes, but also account for the highest numbers of detected SFPs. The bias towards a specific fraction of expressed genes in maize is far less for MF and CF than for mRNA, because 95% of maize exons are unmethylated (Rabinowicz et al., 2003) and CF gene enrichment is independent of methylation and gene expression patterns (Peterson et al., 2002b).

Even when the cRNA or DNA target sequence was identical to the PM probe sequence, we observed instances where the MM probe had higher signal intensity. Possible explanations for this unexpected outcome are as follows: First, the quantity of hybridized target sequence may be low, resulting in a PM probe intensity that is difficult to separate from the overall background noise. Most PM probes ineffective for SFP genotyping with mRNA-derived cRNA are hindered by low gene expression levels. Second, spurious hybridization of sequences with high similarity to the MM probe could have masked the true target signal. Compared to GeneChips hybridized with cRNA, all genomic DNA target fractions presumably have higher amounts of spurious repetitive DNAs diluting the PM signal. Based on a previously published repeat analysis of CF and MF maize genome sequencing data, the total number of repeat sequences in MF and CF libraries was 33% (17,419/52,649) and 14% (10,154/71,492), respectively (Whitelaw et al., 2003). While our CF and MF libraries

did not meet the exact specifications of those analyzed in the above study, these findings indicate that residual repetitive DNAs are almost certainly co-hybridized to CF and MF arrays. In particular, a higher percentage of array probes hybridized with AFLP samples is clearly not useful for scoring SFPs. This is not an unexpected outcome given that the 125 Mb AFLP target fraction has a low percentage of amplified sequences complementary to probe sequences. Whatever the cause, probe pairs for which the target is known to be an exact match to the PM probe and of those that have a large MM/PM signal ratio are most likely ineffective for detecting sequence polymorphisms.

As shown in Table 2.5, another point of interest lies in the fact that the power to detect polymorphic probesets was much greater than the power to detect individual probes at comparable false discovery rates. At least two factors contribute to this difference: First and foremost is the large amount of data available to test probe by line interaction in a probeset. All 135 data points from fifteen probes on nine arrays can be used, whereas a comparison of two lines at a single probe involves only six data points. This discrepancy, however, cannot explain why the gain in power was much greater for the mRNA method than for the DNA methods. A likely explanation is that differences in gene expression levels interfere with the ability to detect probe by line interaction with the mRNA method but not with the DNA methods. We did not take into account varying DNA and gene expression levels when calculating probe intensity differences between lines because we found that doing so resulted in lower power for all methods, even the mRNA method (data not shown).

While CF is broadly applicable to both plants and animals, it is technically challenging to generate reproducible libraries from multiple diverse genotypes and to optimize the method for high-throughput applications. MF, on the other hand, is specific to plants and the level of gene-enrichment is species dependent (Rabinowicz

et al., 2005). Gel purification of the unmethylated gene-rich fraction of plant genomes is also not highly amenable to rapid processing, and cytosine methylation differences between genotypes are known to create non-SNP polymorphisms (Cervera et al., 2002). Moreover, residual genome complexity consisting of repetitive DNA in both CF and MF samples is believed to have complicated SFP detection in this study.

As discussed above, the target preparation methods evaluated in this study offered only modest power to detect SFPs with the Maize GeneChip. The effective use of such arrays for genotyping complex plant genomes would require several improvements, including custom array designs with additional replication and tiling of probes and more aggressive reduction of genomic complexity than can be accomplished via standard MF and CF approaches (e.g., MF, followed by HC). AFLP is expected to be a more powerful method in such cases, provided that probes are selected from sequences represented in the AFLP sample used for hybridization. By using an AFLP design similar to whole-genome sampling analysis (WGSA) in humans (Kennedy et al., 2003), it may be possible to selectively SNP genotype amplified gene fragments and promote reduction of genome complexity to the desired level.

REFERENCES

- Arumuganathan, K., and E.D. Earle. 1991. Nuclear DNA content of some important plant species. *Plant Mol. Biol. Rep.* 9:208-218.
- Bedell, J.A., M.A. Budiman, A. Nunberg, R.W. Citek, D. Robbins, J. Jones, E. Flick, T. Rholffing, J. Fries, K. Bradford, J. McMenamy, M. Smith, H. Holeman, B.A. Roe, G. Wiley, I.F. Korf, P.D. Rabinowicz, N. Lakey, W.R. McCombie, J.A. Jeddelloh, and R.A. Martienssen. 2005. Sorghum genome sequencing by methylation filtration. *PLoS Biol.* 3:e13.
- Bernardi, G. 1971. Chromatography of nucleic acids on hydroxyapatite columns. *Methods Enzymol.* 21: 95-139.
- Bolstad, B.M., R.A. Irizarry, M. Astrand, and T.P. Speed. 2003. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185-193.
- Borevitz, J.O., D. Liang, D. Plouffe, H.S. Chang, T. Zhu, D. Weigel, C.C. Berry, E. Winzeler, and J. Chory. 2003. Large-scale identification of single-feature polymorphisms in complex genomes. *Genome Res.* 13:513-523.
- Britten, R.J., and D.E. Kohne. 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. *Science* 161:529-540.
- Cervera, M.T., L. Ruiz-Garcia, and J.M. Martinez-Zapater. 2002. Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. *Mol. Genet. Genomics* 268:543-552.
- Cui, X., J. Xu, R. Asghar, P. Condamine, J.T. Svensson, S. Wanamaker, N. Stein, M. Roose, and T.J. Close. 2005. Detecting single-feature polymorphisms using oligonucleotide arrays and robustified projection pursuit. *Bioinformatics* 21:3852-3858.
- Fu, H., Z. Zheng, and H.K. Dooner. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl. Acad. Sci.* 99:1082-1087.
- Fu, H., W. Park, X. Yan, Z. Zheng, B. Shen, and H.K. Dooner. 2001. The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. *Proc. Natl. Acad. Sci.* 98:8903-8908.

- Gaut, B.S., M. Le Thierry d'Ennequin, A.S. Peek, and M.C. Sawkins. 2000. Maize as a model for the evolution of plant nuclear genomes. *Proc. Natl. Acad. Sci.* 97:7008-7015.
- Geever, R.F., F.R.H. Katterman, and J.E. Endrizzi. 1989. DNA hybridization analyses of a *Gossypium* allotetraploid and two closely related diploid species. *Theor. Appl. Genet.* 77:553-559.
- Hake, S., and V. Walbot. 1980. The genome of *Zea mays*, its organization and homology to related grasses. *Chromosoma* 79:251-270.
- Hazen, S.P., and S.A. Kay. 2003. Gene arrays are not just for measuring gene expression. *Trends Plant Sci.* 8:413-416.
- Ihaka, R., and R. Gentleman. 1996. R: A language for data analysis and graphics. *J. Comput. Graph. Statist.* 5:299-314.
- Irizarry, R.A., B. Hobbs, F. Collin, Y.D. Beazer-Barclay, K.J. Antonellis, U. Scherf, and T.P. Speed. 2003. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264.
- Kennedy, G.C., H. Matsuzaki, S. Dong, W.M. Liu, J. Huang, G. Liu, X. Su, M. Cao, W. Chen, J. Zhang, W. Liu, G. Yang, X. Di, T. Ryder, Z. He, U. Surti, M.S. Phillips, M.T. Boyce-Jacino, S.P. Fodor, and K.W. Jones. 2003. Large-scale genotyping of complex DNA. *Nat. Biotechnol.* 21:1233-1237.
- Kirst, M., R. Caldo, P. Casati, G. Tanimoto, V. Walbot, R.P. Wise, and E.S. Buckler. 2006. Genetic diversity contribution to errors in short oligonucleotide microarray analysis. *Plant Biotechnol. J.* 4:489-498.
- Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. *Genetics* 165:2117-2128.
- Liu, W.M., R. Mei, X. Di, T.B. Ryder, E. Hubbell, S. Dee, T.A. Webster, C.A. Harrington, M.H. Ho, J. Baid, and S.P. Smeeckens. 2002. Analysis of high density expression microarrays with signed-rank call algorithms. *Bioinformatics* 18:1593-1599.
- Palmer, L.E., P.D. Rabinowicz, A.L. O'Shaughnessy, V.S. Balija, L.U. Nascimento, S. Dike, M. de la Bastide, R.A. Martienssen, and W.R. McCombie. 2003. Maize genome sequencing by methylation filtration. *Science* 302:2115-2117.

- Peterson, D.G., S.R. Schulze, E.B. Sciara, S.A. Lee, J.E. Bowers, A. Nagel, N. Jiang, D.C. Tibbitts, S.R. Wessler, and A.H. Paterson. 2002a. Integration of Cot analysis, DNA cloning, and high-throughput sequencing facilitates genome characterization and gene discovery. *Genome Res.* 12:795-807.
- Peterson, D.G., S.R. Wessler, and A.H. Paterson. 2002b. Efficient capture of unique sequences from eukaryotic genomes. *Trends Genet.* 18:547-550.
- Rabinowicz, P.D., L.E. Palmer, B.P. May, M.T. Hemann, S.W. Lowe, W.R. McCombie, and R.A. Martienssen. 2003. Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res.* 13:2658-2664.
- Rabinowicz, P.D., K. Schutz, N. Dedhia, C. Yordan, L.D. Parnell, L. Stein, W.R. McCombie, and R.A. Martienssen. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* 23:305-308.
- Rabinowicz, P.D., R. Citek, M.A. Budiman, A. Nunberg, J.A. Bedell, N. Lakey, A.L. O'Shaughnessy, L.U. Nascimento, W.R. McCombie, and R.A. Martienssen. 2005. Differential methylation of genes and repeats in land plants. *Genome Res.* 15:1431-1440.
- Raleigh, E.A. 1992. Organization and function of the *mcrBC* genes of *Escherichia coli* K-12. *Mol. Microbiol.* 6:1079-1086.
- Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler IV. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci.* 98:11479-11484.
- Ronald, J., J.M. Akey, J. Whittle, E.N. Smith, G. Yvert, and L. Kruglyak. 2005. Simultaneous genotyping, gene-expression measurement, and detection of allele-specific expression with oligonucleotide arrays. *Genome Res.* 15:284-291.
- Rostoks, N., J.O. Borevitz, P.E. Hedley, J. Russell, S. Mudie, J. Morris, L. Cardle, D.F. Marshall, and R. Waugh. 2005. Single-feature polymorphism discovery in the barley transcriptome. *Genome Biol.* 6:R54.
- Saghai-Maroo, M.A., K.M. Soliman, R.A. Jorgensen, and R.W. Allard. 1984. Ribosomal DNA spacer-length polymorphisms in barley Mendelian inheritance chromosomal location and population dynamics. *Proc. Natl. Acad. Sci.* 81:8014-8018.

- SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melakeberhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765-768.
- Singer, T., Y. Fan, H.-S. Chang, T. Zhu, S. P. Hazen, and S.P. Briggs. 2006. A High-Resolution Map of *Arabidopsis* Recombinant Inbred Lines by Whole-Genome Exon Array Hybridization. *PLoS Genetics* 2:e144.
- Steinmetz, L.M., H. Sinha, D.R. Richards, J.I. Spiegelman, P.J. Oefner, J.H. McCusker, and R.W. Davis. 2002. Dissecting the architecture of a quantitative trait locus in yeast. *Nature* 416:326-330.
- Stupar, R.M. and N.M. Springer. 2006. *Cis*-transcriptional variation in maize inbred lines B73 and Mo17 leads to additive expression patterns in the F₁ hybrid. *Genetics* 173: 2199–2210.
- Sutherland, E., L. Coe, and E.A. Raleigh. 1992. McrBC: a multisubunit GTP-dependent restriction endonuclease. *J. Mol. Biol.* 225:327.
- Syvänen, A.C. 2005. Toward genome-wide SNP genotyping. *Nat. Genet.* 37 Suppl.:S5-10.
- Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci.* 98:9161-9166.
- Tsolaki, A.G., A.E. Hirsh, K. DeRiemer, J.A. Enciso, M.Z. Wong, M. Hannan, Y.O. Goguet de la Salmoniere, K. Aman, M. Kato-Maeda, and P.M. Small. 2004. Functional and evolutionary genomics of *Mycobacterium tuberculosis*: insights from genomic deletions in 100 strains. *Proc. Natl. Acad. Sci.* 101:4865-4870.
- Vos, P., R. Hogers, M. Bleeker, M. Reijans, T. van de Lee, M. Hornes, A. Frijters, J. Pot, J. Peleman, M. Kuiper, and et al. 1995. AFLP: a new technique for DNA fingerprinting. *Nucleic Acids Res.* 23:4407-4414.
- Werner, J.D., J.O. Borevitz, N. Warthmann, G.T. Trainer, J.R. Ecker, J. Chory, and D. Weigel. 2005. Quantitative trait locus mapping and DNA array hybridization identify an FLM deletion as a cause for natural flowering-time variation. *Proc. Natl. Acad. Sci.* 102:2460-2465.
- White, S.E., and J.F. Doebley. 1999. The molecular evolution of terminal ear1, a regulatory gene in the genus *Zea*. *Genetics* 153:1455-1462.

- Whitelaw, C.A., W.B. Barbazuk, G. Pertea, A.P. Chan, F. Cheung, Y. Lee, L. Zheng, S. van Heeringen, S. Karamycheva, J.L. Bennetzen, P. SanMiguel, N. Lakey, J. Bedell, Y. Yuan, M.A. Budiman, A. Resnick, S. Van Aken, T. Utterback, S. Riedmuller, M. Williams, T. Feldblyum, K. Schubert, R. Beachy, C.M. Fraser, and J. Quackenbush. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302:2118-2120.
- Winzeler, E.A., D.R. Richards, A.R. Conway, A.L. Goldstein, S. Kalman, M.J. McCullough, J.H. McCusker, D.A. Stevens, L. Wodicka, D.J. Lockhart, and R.W. Davis. 1998. Direct allelic variation scanning of the yeast genome. *Science* 281:1194-1197.
- Wolyn, D.J., J.O. Borevitz, O. Loudet, C. Schwartz, J. Maloof, J.R. Ecker, C.C. Berry, and J. Chory. 2004. Light-response quantitative trait loci Identified with composite interval and eXtreme array mapping in *Arabidopsis thaliana*. *Genetics* 167:907-917.
- Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeu, B.J. Nikolau, and P.S. Schnable. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. *Proc. Natl. Acad. Sci.* 99:6157-6162.
- Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High-Cot sequence analysis of the maize genome. *Plant J.* 34:249-255.
- Zhao, W., P. Canaran, R. Jurkuta, T. Fulton, J. Glaubitz, E. Buckler, J. Doebley, B. Gaut, M. Goodman, J. Holland, S. Kresovich, M. McMullen, L. Stein, and D. Ware. 2006. Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.* 34:D752-757.
- Zhou, Y., T. Bui, L.D. Auckland, and C.G. Williams. 2002. Undermethylated DNA as a source of microsatellites from a conifer genome. *Genome* 45:91-99.

CHAPTER 3

LARGE-SCALE ENRICHMENT AND DISCOVERY OF GENE-ENRICHED SNPS³

ABSTRACT

Whole-genome association studies of complex traits in higher eukaryotes require a high density of single nucleotide polymorphism (SNP) markers at genome-wide coverage. To design high-throughput, multiplexed SNP genotyping assays, researchers must first discover large numbers of SNPs by extensively resequencing multiple individuals or lines. For SNP discovery approaches using short read lengths that next-generation DNA sequencing technologies offer, the highly repetitive and duplicated nature of large plant genomes presents additional challenges. Here, we describe a genomic library construction procedure that facilitates pyrosequencing of genic and low-copy regions in plant genomes, and a customized computational pipeline to analyze and assemble short reads (100-200 bp), identify allelic reference sequence comparisons, and call SNPs with a high degree of accuracy. With maize (*Zea mays* L.) as the test organism in a pilot experiment, the implementation of these methods resulted in the identification of 126,683 putative SNPs between two maize inbred lines at an estimated false discovery rate (FDR) of 15.1%. We estimated rates of false SNP discovery using an internal control, and we validated these FDR rates with an external SNP dataset that was generated using locus specific PCR amplification and Sanger sequencing. These results show that this approach has wide applicability for efficiently and accurately detecting gene-enriched SNPs in large,

³ M. A. Gore, M. H. Wright, E. S. Ersoz, P. Bouffard, E. S. Szekeres, T. P. Jarvie, B. L. Hurwitz, A. Narechania, T. T. Harkins, G. S. Grills, D. H. Ware, E. S. Buckler. 2009. Large-scale enrichment and discovery of gene-enriched SNPs. *The Plant Genome*. In press

complex plant genomes.

INTRODUCTION

The average nucleotide diversity of coding regions between any two maize lines ($\pi=1-1.4\%$) is 2- to 5-fold higher than other domesticated grass crops (Buckler et al., 2001; Tenaillon et al., 2001; Wright et al., 2005). Moreover, it is not uncommon to find maize haplotypes more than 2% diverged from one another (Tenaillon et al., 2001; Wright et al., 2005) and even as high as 5% (Henry and Damerval, 1997). Intragenic linkage disequilibrium (LD) rates rapidly decline to nominal levels within 2 kb in a population of diverse maize inbred lines (Remington et al., 2001). Of the ~2500 Mb that constitutes the maize genome, less than 25% is genic or low-copy-number sequence, with large blocks of highly repetitive DNA such as retrotransposons intermixed throughout (Hake and Walbot, 1980; Meyers et al., 2001; SanMiguel et al., 1996). Retrotransposons are generally recombinationally inert, and most meiotic recombination in the maize genome is restricted to gene-rich regions (Fu et al., 2002; Fu et al., 2001; Yao et al., 2002). Association mapping strategies, which rely on ancient recombination for dissecting complex traits, require that SNPs within these recombinationally active gene regions be identified and genotyped in phenotypically diverse populations (Reviewed by Zhu et al., 2008). Because of the rapid decay of intragenic LD in a highly diverse genome with an estimated 59,000 genes (Messing et al., 2004), several million gene-enriched SNP markers may be necessary for whole-genome association studies in diverse maize (E. Buckler, unpublished).

Retrotransposons contain a higher density of methylation in the form of 5-methylcytosine relative to genic sequences—a property unique to plant genomes (Rabinowicz et al., 2003; Rabinowicz et al., 2005). HypoMethylated Partial Restriction (HMPR) is a library construction method that exploits this property to

facilitate the efficient sequencing of gene rich regions in large, highly repetitive plant genomes (Emberton et al., 2005). The principle underlying HMPR is that the complete digestion of plant genomic DNA with a 5-methylcytosine-sensitive (MCS) restriction enzyme that has a 4 bp recognition sequence permits the fractionation of genic and repetitive DNA by gel electrophoresis. Large restriction fragments (20-150 kb) contain blocks of highly methylated retrotransposons, while much smaller fragments (<1000 bp) comprise a fraction that is gene-enriched (Bennetzen et al., 1994; Yuan et al., 2002). Emberton et al. (2005) used a partial digestion of maize genomic DNA with a MCS 4 bp cutter, followed by gel-purification and cloning procedures to construct maize HMPR libraries that contained larger (1-4 kb), overlapping gene fragments more suitable for Sanger sequencing read lengths (800-1200 bases). These maize HMPR libraries showed more than 6-fold enrichment for genes compared to control libraries. This level of gene enrichment was comparable to that achieved by other non-transcriptome-based gene-enrichment sequencing technologies tested on maize (Gore et al., 2007; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003), but maize HMPR libraries were superior for repeat elimination and enrichment of low-copy, non-coding sequences.

With the recent emergence of ‘next-generation’ DNA sequencing technologies it is technically feasible to economically and rapidly resequence hundreds of millions of bases (Reviewed by Mardis, 2008). Using these high-throughput sequencing-by-synthesis (Bennett, 2004; Margulies et al., 2005) or sequencing-by-ligation (Shendure et al., 2005) technologies in a read-to-reference based SNP discovery approach presents computational challenges because the length and quality of obtained individual reads are shorter and potentially of lower fidelity than single-pass Sanger sequencing reads. Furthermore, the maize genome is the product of ancient and perhaps more recent tetraploidization and rearrangement events (Gaut and Doebley,

1997; Swigoňová et al., 2004; Wei et al., 2007), and as a result contains a high proportion of duplicated genes (Blanc and Wolfe, 2004; Emrich et al., 2007; Messing et al., 2004). This confounds the unique mapping of short reads if duplicated genes (i.e., paralogs) are recently diverged and thus nearly identical in nucleotide sequence. Recently, a computational SNP calling pipeline built on the POLYBAYES polymorphism detection software (Marth et al., 1999) and “monoallelism” rules was developed and used to analyze expressed sequence tags (ESTs) that were obtained by 454 pyrosequencing of cDNAs prepared from two maize inbred lines (Barbazuk et al., 2007). This pipeline reduced the number of false positive SNPs that resulted from sequencing errors and alignment of paralogous sequences, which facilitated the identification of more than 7,000 putative SNPs in expressed genes.

Nonetheless, if the discovery of maize SNP markers on the order of millions is to be economically viable, the use of low cost, next-generation DNA sequencing technologies is clearly required. These high-throughput DNA sequencing technologies can be more efficiently used in the large-scale discovery of SNPs for maize association mapping studies if resequencing is concentrated within the recombinationally active gene regions of the vastly repetitive maize genome. The objectives of this study were (i) to adapt HMPR gene-enrichment sequencing to a massively parallel pyrosequencing platform and (ii) to develop a read-to-reference based SNP calling pipeline for short reads (100-200 bp) that maximizes SNP detection power, while controlling the number of detected false positive SNPs resulting from sequencing errors and the alignment of paralogous sequences.

MATERIALS AND METHODS

DNA Isolation from Maize

We extracted nuclear DNA from nuclei prepared from etiolated (pale green),

inner husk leaves (100 g) of field-grown maize inbred line B73 as previously described by Rabinowicz (2003).

A more specialized cultivation technique was required to obtain genomic DNA from maize root tissue. Kernels from maize inbred lines B73 and Mo17 were surface sterilized in a 10% (vol/vol) bleach solution (5.25% Sodium Hypochlorite) by gently rocking for 30 min, followed by 3X 10 min rinses with sterile water. The kernels were left to imbibe overnight in sterile water at room temperature with gentle rocking. Ten kernels were placed in a vertically orientated seed germination pouch (Mega International, West St. Paul, MN) and germinated in a dark growth chamber held at 28°C. Roots of 1-wk-old maize seedlings were bulk harvested and immediately frozen in liquid N₂ prior to storage at –80°C. Total genomic DNA was isolated from homogenized frozen 1-week-old root tissue using the DNeasy Plant Maxi Kit (QIAGEN, Valencia, CA) according to the manufacturer's protocol.

Modified HMPR Library Construction

Complete digestions of 5 µg of maize husk nuclear DNA (B73) and seedling root total genomic DNA (B73 and Mo17) were individually performed in 100 µL volumes with 50 U of *Hpa*II (New England Biolabs, Ipswich, MA) at 37°C for 16 h, followed by heat inactivation of the enzyme at 65°C for 20 min. *Hpa*II fragments ranging in size from >10 kb to less than 100 bp (data not shown) were separated on a low melting 0.8% SeaPlaque agarose gel (Cambrex Bio Science Rockland, Inc., Rockland, ME). Restriction fragments ranging in size from 100–600 bp were excised from the gel and purified using the QIAquick Gel Extraction kit (QIAGEN, Valencia, CA), according to the manufacturer's protocol. Gel-isolated *Hpa*II fragments were randomly ligated to each other with 1 µL of highly concentrated T4 DNA ligase (20 U/µL) (New England Biolabs, Ipswich, MA) in a total reaction volume of 20 µl at

16°C for 16 h, followed by heat inactivation of the enzyme at 65°C for 20 min.

Several micrograms of concatenated *HpaII* fragments were needed for the downstream nebulization procedure (see 454 sequencing and data processing section). However, this would typically require low-throughput, large-scale DNA extractions and gel isolations, because an estimated 95% of the maize genome was intentionally discarded. Alternatively, we found it more efficient to generate microgram quantities of concatenated *HpaII* fragments using Phi29-based isothermal amplification of long concatemer templates in a nanogram-scale reaction. Briefly, the GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Piscataway, NJ) was used to amplify 1 µL of the 10 ng/µL ligation reaction per the manufacturer's instructions. This kit uses the high fidelity Phi29 (φ29) DNA polymerase, dNTPs, and random hexamers to replicate linear genomic DNA by multiple displacement amplification. Several independent GenomiPhi amplification reactions were performed and pooled for each library to ensure a low level of amplification-induced bias. The GenomiPhi reaction was separated on a low melting 0.8% SeaPlaque Agarose gel, and amplification products ranging in size from 3-10 kb were isolated from the gel with the QIAquick Gel Extraction kit and used in the downstream 454 sample preparation procedure.

454 Sequencing and Data Processing

Sequence sample preparation and data generation were performed with the Phi29 amplified *HpaII* concatemer DNA of two B73 HMPR libraries (husk and root) and one Mo17 HMPR library (root) using the 454 GS FLX platform at 454 Life Sciences (Branford, CT). In addition, total genomic DNA isolated from the same seedling root tissue of B73 was sequenced on the same 454 platform, which served as an unfiltered (UF) genomic control to assess the level of gene-enrichment in modified HMPR libraries. Approximately 5 µg of high molecular weight DNA was fragmented

by nebulization to a size range of 300–500 bp. Preparation of 454 libraries, emulsion-based clonal amplification, library sequencing on the Genome Sequencer FLX System as well as signal processing and data analysis were performed as previously described by Margulies et al. (2005). Also, the 454 base-calling software (version 1.1.03.24) provided error estimates (Q values) for each base, none of which exceeded a value of 40.

The expected yield per run of the 454 GS FLX is approximately 100 Mb, potentially more under ideal conditions. However, sequencing the B73 husk library with a single instrument run produced only 65.6 Mb of sequence because a less than optimal DNA copy per bead ratio was used for emulsion PCR. A more optimal DNA copy per bead ratio was used for the B73 root library, improving sequence yield to 101.3 Mb in a single run. The Mo17 root library was sequenced with four runs that in total yielded 236.7 Mb of sequence. This total sequence yield for the Mo17 root library was 41% lower than expected, indicating that further optimization was still needed. In addition, we sequenced (1 run; 130.9 Mb) randomly sheared B73 total genomic DNA, which served as the UF library.

The raw 454 sequencing data are available in the NCBI Short Read Archive (<http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi>).

Screening and Filtering of 454 Sequences

Because modified HMPR libraries contained *HpaII* concatemers, 454 reads generated from sequencing these libraries were digested *in silico* at *HpaII* recognition sites (5'–C/CGG–3'). This was done to produce independent, non-chimeric *HpaII* fragment sequences. All 454 reads from the UF control library and *HpaII* fragment sequences less than 40 bp in length were discarded. *HpaII* fragment sequences and UF sequences (≥ 40 bp) were searched using BLAT (Kent, 2002) against The Institute for

Genomic Research (TIGR) maize repeat database Version 4.0 (http://maize.tigr.org/repeat_db.shtml) to identify repetitive sequences. Also, sequences were searched against mitochondrial (GenBank accession no. NC_007982.1) and chloroplast (GenBank accession no. NC_001666.2) genome sequences of maize. We performed BLAT searches with default parameters, except for a tile size of 16. We considered BLAT similarities significant if the expectation value was less than 10^{-5} and the local alignment length was 40 bp or longer. Sequences that had a significant match to a repeat sequence or an organellar genome were discarded. Remaining sequences were similarly searched with BLAT against the Maize Assembled Genome Island Version 4.0 Contigs and Singletons (MAGIv4.0 C&G) database (<http://magi.plantgenomics.iastate.edu/>). Because a large number of sequences did not match any sequences in the MAGIv4.0 C&G database, these unmatched *Hpa*II fragment and UF sequences were also searched against the complete genome sequences of *japonica* rice (*Oryza sativa* L.) (<http://rice.plantbiology.msu.edu/>) and sorghum (*Sorghum bicolor* L.) (<http://www.phytozome.net/>) as well as maize expressed sequence tag (EST) sequences within the Dana-Farber Cancer Institute (DFCI) maize gene index release 17.0 (<http://compbio.dfci.harvard.edu/tgi/>). Sequences that did not have a significant match in any of these additionally searched databases were considered contaminant (non-maize) sequence and discarded. Summary statistics and source information for all databases are found in Table 3.1.

Table 3.1. Statistics of databases and genome sequences used in this study.

Name	Source	Sequences	Mb
Screening Databases			
Maize Chloroplast Genome	GenBank acc. no. NC_001666.2	1	0.140
Maize Mitochondrial Genome	GenBank acc. no. NC_007982.1	1	0.569
TIGR Maize Repeatv4.0	http://maize.tigr.org/repeat_db.shtml	26,791	19.6
Rice	http://rice.plantbiology.msu.edu/	12	372.1
Sorghum	http://www.phytozome.net/	10	697.6
DFCI Maize Gene Index release 17.0	http://compbio.dfci.harvard.edu/tgi/	115,744	86.3
Reference Database			
MAGIv4.0 Contigs and Singletons	http://magi.plantgenomics.iastate.edu/	727,781	675.2

Assembly of 454 Sequences

We assembled the retained non-repeat *HpaII* fragment sequences into multiple sequence alignments using the CAP3 sequence assembly program (Huang and Madan, 1999). The following CAP3 assembly options were used: -p 99 (overlaps must be >99% identity), -s 401 (alignment score must be >400, minimum value allowed), -h 3 (maximum overhang of 3%), and alignment scoring options (-m 20, -n 40, and -g 21) that allowed a perfect match overlap of 40 bp to satisfy the minimum alignment score for assembly. Additionally, CAP3 computed a *Q* value for each base of the consensus sequence. Assemblies were performed separately for B73 (husk and root) and Mo17 (root) non-repeat *HpaII* fragment sequences. We did not assemble UF sequences, as they were only used to measure the level of gene-enrichment and repeat depletion in modified HMPR libraries.

Because CAP3 could not execute with all sequences input at once, we performed a preliminary clustering of sequences into a collection of disjoint groups with no inter-group homology. Clustering was performed by a custom program in a manner equivalent to NCBI BLASTClust (available at <http://www.ncbi.nlm.nih.gov/BLAST/docs/blastclust.html>). We did not use BLASTClust because it could not run on our systems with the amount of input data

supplied. CAP3 was then executed on each cluster separately. The preliminary clustering revealed that about 5% of sequences were still chimeric because of an *HpaII* site that was eliminated by a sequencing error or erroneous end-joining ligation. A simple modification to the clustering algorithm allowed almost all chimeras to be detected and split before CAP3 assembly.

We developed a custom program to analyze the CAP3 assembly output and extract a consensus sequence and associated CAP3-based Q values from each multiple sequence assembly as well as the number of sequences concordant with each consensus base (coverage depth). Because of partial overlaps and potential disagreements among assembled reads, coverage depth as defined here is not the same as the total number of reads aligned in the multiple sequence assembly but as the number of reads with an aligned base that supports the consensus base call. *HpaII* fragment sequences that did not assemble into multiple sequence alignments (i.e., singletons) were used directly as consensus sequences as well as the Q values calculated by Roche-454's base-calling software.

Construction of the Paralog Distinguishing List (PDL)

To facilitate the identification of paralogous regions, the MAGIv4.0 C&G database of B73 reference sequences was searched and aligned against itself using BLAT, as described above. All match pairs (not the alignment) with at least 90% identity and a length of 50 bp or longer were used as input for a custom polymorphism detection program. The custom polymorphism detection program performed a Smith-Waterman (Smith and Waterman, 1981) local alignment between match pairs identified by BLAT to obtain a full representation of the alignment “in memory.” This allowed alignments to be quickly scanned for single base mismatches and single base insertions/deletions (in/dels). Single base mismatches and single base in/dels were

identified in the Smith-Waterman local alignments and “context sequences” were extracted: the 16 bp 5' and 16 bp 3' flanking the mismatch or in/del. All such putative non-allelic differences were extracted as context sequences from all pairwise matches satisfying the 90% identity minimum and 50 bp minimum. These context sequences form the PDL and represent the putative fixed differences that distinguish paralogs. The PDL was used in further analysis to search for paralogous regions, as described below.

Polymorphism Detection

Consensus sequences of B73 and Mo17 *HpaII* fragments were searched against B73 reference sequences (MAGIv4.0 C&G database) using BLAT. Match pairs (not the alignments) were used as input for the custom polymorphism detection program, as described above. Similarly, the polymorphism detection program performed a Smith-Waterman local alignment between the *HpaII* consensus sequence and the MAGIv4.0 C&G reference sequence (i.e., match pairs) identified by BLAT to obtain a full representation of the alignment “in memory.” For each single base mismatch or in/del identified by the program, context sequences for B73 and Mo17 *HpaII* fragment sequences were extracted: the 16 bp 5' and 16 bp 3' flanking the mismatch or in/del. Single base mismatches or in/dels within 16 bp of either end of the local alignment were not considered.

Implementation of the Paralog Distinguishing List (PDL) and SNP Calling

With the same custom polymorphism detection program, all context sequences for B73 or Mo17 *HpaII* fragment sequences were searched against the PDL. Any match to the PDL was considered a paralogous alignment and the entire alignment and all potential SNPs within it were discarded. Otherwise, if no PDL matches were found,

all in/del contexts were discarded (not called as SNPs) and the remaining single base mismatch contexts were scanned against a list of SNPs already called. If a single duplicate context was identified in an alignment, only that context was discarded, but if two or more duplicates were identified, the entire alignment was discarded, along with all potential SNPs, even if these SNPs were novel. Provided neither the PDL nor the duplicate alignment check resulted in discarding all potential SNPs, the remaining single base mismatches were called SNPs and no further alignments for the current *HpaII* consensus sequence were considered. Otherwise, if the alignment was discarded, the next strongest BLAT match was considered, continuing until an alignment was accepted, or until the next strongest BLAT match was less than 95% identity. This preset 5% maximum was not restrictive for identifying allelic variation, as it is well above the average nucleotide diversity of coding regions between any two maize lines ($\pi=1-1.4\%$) (Tenailon et al., 2001; Wright et al., 2005), but still allows the evaluation of haplotypes that are 5% diverged from one another (Henry and Damerval, 1997). Moreover, the 5% maximum allowed us to use a smaller PDL by avoiding paralogous alignments that were more diverged and easily distinguished from previously reported allelic variation levels. Identified B73/Mo17 putative SNPs and the PDL are available for download from Panzea (<http://www.panzea.org>).

Panzea SNP Comparison

We extracted 6,094 B73 and 6,200 Mo17 sequences from the Panzea database (Zhao et al., 2006) that were generated by PCR-directed Sanger sequencing of candidate gene loci. Overlapping sequences that were amplified from the same candidate gene locus were assembled using the procedure described above, except that sequences were clustered based on a common Panzea locus ID. For many of the candidate gene loci, there were two independent amplifications and sequencings of

B73 and Mo17 for quality control. This resulted in 3,683 (1.57 Mb) and 3,696 (1.57 Mb) assemblies for B73 and Mo17, respectively. We called SNPs from these sequences using the program already described, except allelic B73 and Mo17 consensus sequences were paired on the basis of common Panzea locus ID. The PDL was not used to call SNPs with Panzea sequences, because it was assumed that all Mo17/B73 pairings were allelic on the basis of single locus PCR amplification. Identified Panzea SNPs were mapped to Mo17 454 consensus sequences on the basis of the 16 bp 5' and 16 bp 3' context sequences, and vice versa, to identify which SNPs from each dataset were called from sequence in common to both datasets. We separately looked at the intersection of Panzea SNPs and B73/Mo17 *HpaII* SNPs called with (126,683 SNPs; no thresholds) and without (174,476 SNPs; no thresholds) the PDL. We then compared SNPs that mapped to both datasets to estimate the rate of false SNP discovery and power, assuming that all true Mo17/B73 SNPs were discovered in the Panzea dataset and no false SNPs were discovered.

All custom code and scripts used in this study are available upon request from M. H. Wright (mhw6@cornell.edu).

RESULTS

Construction of Modified HMPR Libraries

We modified the previously described HMPR library construction method (Emberton et al., 2005) to allow high-throughput gene-enrichment sequencing of the maize genome using the 454 Genome Sequencer FLX (GS FLX) pyrosequencing instrument (see “Materials and Methods”). *HpaII*, a MCS 4 bp cutter (5'-C/CGG-3'), was selected to construct modified HMPR libraries, because of its strong bias for cleaving within unmethylated genic and low-copy regions of the maize genome (Antequera and Bird, 1988; Emberton et al., 2005; Yuan et al., 2002). The first of the

two major modifications to the HMPR method was to allow maize genomic DNA to be completely digested with *HpaII* rather than partially digested. This was done to produce a more repeatable *HpaII* restriction pattern and, as a result, consistently enrich for gene fragments mostly smaller than 600 bp. Second, *HpaII* fragments between the sizes of 100–600 bp were gel-isolated and converted via random ligation into concatemers of longer lengths more suitable for nebulization (i.e., fragmentation). At the time of this experiment, it was not possible for us to execute paired-end read sequencing and to routinely obtain read lengths longer than 250 bases on the 454 GS FLX instrument; thus, we used ligation and nebulization in combination to construct and randomly break *HpaII* concatemers in order to completely sequence larger *HpaII* fragments.

To test and optimize our library construction method, we constructed modified HMPR libraries for maize inbred lines B73 (husk and root) and Mo17 (root). One concern with modified HMPR and its predecessor is the potential enrichment of organellar genome fragments in constructed libraries (Emberton et al., 2005), as these genomes are unmethylated (Palmer et al., 2003) and, depending on the tissue type, may be present at a very high copy number (Li et al., 2006). Thus, we evaluated as sources of genomic DNA two etiolated tissue types that were expected to have a relatively low abundance of chloroplasts: inner husk leaves (pale green) and dark-grown seedling roots (white). For inner husk leaves, purification of nuclei prior to genomic DNA extraction was used to further limit the amount of co-isolated chloroplast DNA. For dark-grown seedling roots, we used a higher yielding and less laborious total genomic DNA extraction procedure that lacked a nuclei purification step, because dark-grown seedling roots were expected to be highly deficient in chloroplasts and other types of plastids (Reviewed by Possingham, 1980).

Compositional Analysis of Modified HMPR Libraries

Modified HMPR libraries and an unfiltered (UF) B73 library were sequenced on the 454 GS FLX instrument (see “Materials and Methods”). Because the modified HMPR libraries were comprised of randomly concatenated *HpaII* fragments (see previous section), prior to analysis 454 reads pertaining to these libraries were *in silico* digested with *HpaII* to produce independent, non-chimeric sequences. To examine the sequence composition of modified HMPR and UF libraries, *HpaII* fragment and UF sequences were searched against several plant nucleotide databases and genome sequences (see “Materials and Methods”). The distribution of sequence among these categories is shown in Table 3.2. A higher level of organellar contamination was found in root libraries, but this was offset by their lower level of repeats. B73 and Mo17 root libraries were 7- to 8-fold lower in repeats relative to the B73 husk library, and 14- to 16-fold lower in repeats relative to the UF library. The very low repeat content of root libraries is comparable to that previously reported in maize HMPR libraries (Emberton et al., 2005) and superior to other non-transcriptome-based gene-enrichment sequencing technologies tested on maize (Gore et al., 2007; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003). Even though the amount of repeat sequences within modified HMPR libraries varied substantially between tissue types (e.g., B73 husk vs. B73 root), additional biological and technical replications are needed to determine if these differences are attributed to tissue-specific differential methylation of genes and repeats.

Table 3.2. Sequence composition of modified HMPR and UF libraries.

Libraries	Modified HMPR											
	B73 Husk			B73 Root			Mo17 Root			UF		
	No.	Mb	% [†]	No.	Mb	%	No.	Mb	%	No.	Mb	%
454 reads [‡]	391,778	65.6	-	470,918	101.2	-	1,284,692	236.7	-	543,385	130.9	-
Total [§]	479,565	63.6	100	771,557	97.6	100	1,937,032	225.5	100	543,350	130.9	100
Chloroplast	3,771	0.6	0.8	5,567	0.9	0.7	30,835	4.1	1.6	3,118	0.8	0.6
Mitochondrial	1,319	0.2	0.3	20,332	3.0	2.6	224,593	29.7	11.6	5,493	1.4	1.0
Non-maize [¶]	6,829	0.9	1.4	530,876	67.4	68.8	454,413	49.1	23.5	41,149	9.8	7.6
Repeats [#]	150,786	21.7	31.4	34,378	5.2	4.5	75,225	9.3	3.9	343,072	83.8	63.1
Non-repeats ^{††}	316,860	40.2	66.1	180,404	21.1	23.4	1,151,966	133.3	59.5	150,518	35.1	27.7

[†]The number of sequences in each category expressed as a percentage of the total number of sequences.

[‡]Sequencing reads generated on the 454 GS FLX.

[§]454 reads from modified HMPR libraries were *in silico* digested with *HpaII*, and only sequences ≥ 40 bp were kept and BLAT searched against nucleotide databases. 454 reads from the UF library were not *in silico* digested with *HpaII*, and only sequences ≥ 40 bp were kept and BLAT searched against nucleotide databases.

[¶]Sequences that did not significantly match any of the screened plant nucleotide, organellar, or repeat databases. All of these sequences were classified as putatively non-maize with the majority of unknown or bacterial origin.

[#]Sequences from the maize nuclear genome that significantly matched to The Institute for Genomic Research (TIGR) Maize Repeat version 4 database, which consists of characterized, uncharacterized, and predicted repeats.

^{††}Sequences from putatively non-repetitive regions of the maize genome with significant matches to the Maize Assembled Gene Islands Version 4.0 Contigs and Singletons (MAGIv4.0 C&S) database, sorghum or rice genome sequences, or the Dana-Farber Cancer Institute (DFCI) maize gene index.

The desired enrichment for the genic fraction of the maize genome in root libraries was compromised by an abundance of sequences that did not significantly match any of the screened plant nucleotide databases or genome sequences. These unknown contaminant sequences were most prevalent in the B73 root library, comprising 68.8% of the *HpaII* fragment sequences. We randomly sampled 1,000 of these putative non-maize sequences from each root library and searched them with BLAST (Altschul et al., 1997) against NCBI's non-redundant nucleotide database. On average, 65% of these sampled sequences had no significant similarity (cutoff E-value of 10^{-5}) to any sequence with another 30% showing different degrees of similarity to bacterial sequences (results not shown). We suspect that bacterial endo- or exo-symbionts of maize roots were living beneath the seed pericarp layer and subsequently proliferated on seedling roots. Neither the seed surface sterilization procedure nor the sterile seedling growth conditions used in this study would have eliminated any type of bacterial symbiont from seedling roots, thus allowing the co-isolation of bacterial genomic DNA and its enrichment in modified HMPR root libraries. Regardless of the source or identity of these sequences, these putatively non-maize sequences as well as the maize repeat and organellar sequences were excluded from further analyses.

To assess the degree to which modified HMPR libraries were enriched with genic sequences, we searched non-repetitive, maize *HpaII* sequences against the Maize Assembled Genome Island version 4.0 Contigs and Singletons (MAGIv4.0 C&S) database (<http://magi.plantgenomics.iastate.edu/>). The MAGIv4.0 C&S database is a partial genome assembly of Sanger-based BAC end and shotgun sequences, gene-enriched genome survey sequences as well as whole-genome shotgun sequences from maize inbred line B73 (Kalyanaraman et al., 2007). In addition, the MAGIv4.0 C&S database represents the most comprehensive maize genomic database

in advance of the pending draft maize genome sequence.⁴ The search results revealed an intermediate to high intersection (52.2–67.0%) between the MAGIv4.0 C&S database and non-repetitive *Hpa*II fragment sequences contained within modified HMPR libraries (Table 3.3). Moreover, alignment to computationally predicted genes from MAGIv4.0 Contig sequences and the Dana-Farber Cancer Institute (DFCI) maize gene index (<http://compbio.dfci.harvard.edu/tgi/>) showed that modified HMPR libraries were 4- to 5-fold enriched for genes relative to the UF library (Table 3.3). This level of gene-enrichment in modified HMPR libraries was similar to that obtained with the original HMPR method (Emberton et al., 2005) and other non-EST-based gene-enrichment sequencing technologies tested on maize (Gore et al., 2007; Palmer et al., 2003; Rabinowicz et al., 1999; Whitelaw et al., 2003; Yuan et al., 2003).

Sequence Assembly and Construction of a Paralog Distinguishing List (PDL)

Why is it challenging to identify SNPs in maize using next generation sequencing technologies? Maize is hypothesized to be an ancient tetraploid (Gaut and Doebley, 1997; Swigoňová et al., 2004; Wei et al., 2007), but its genome has lost a substantial number of unlinked duplicated genes (Lai et al., 2004). However, nearly one-third of all maize genes still have a paralog (Blanc and Wolfe, 2004), and many of these paralogs are tandemly arrayed (Messing et al., 2004). It is estimated, based on ESTs, that maize paralogs resulting from an ancient tetraploid event have diverged a minimum of 10% over time (Blanc and Wolfe, 2004), but recent evidence conservatively suggests that nearly identical paralogs ($\geq 98\%$ identity) are almost 13-fold more frequent in the maize genome than that of *Arabidopsis* (Emrich et al., 2007). With long enough sequencing reads, unique flanking sequence can be found to

⁴ The unassembled, draft maize B73 genome sequence is a superior reference sequence, but its use in this study was restricted by the Ft. Lauderdale agreement governing the pre-publication use of large genomic datasets.

Table 3.3 Gene Enrichment Analysis of modified HMPR and UF libraries.

	Modified HMPR						UF	
	B73 Husk		B73 Root		Mo17 Root			
Databases	No.	% [†]	No.	%	No.	%	No.	%
MAGIv4.0 contigs and singletons [‡]	244,189	52.2	131,398	61.2	822,117	67.0	124,323	25.2
MAGIv4.0 contigs [§]	207,576	44.4	118,367	55.1	784,094	61.0	87,387	17.7
MAGIv4.0 contigs genes [¶]	129,095	27.6	75,453	35.1	501,116	40.8	41,004	8.3
DFCI maize gene index [#]	75,027	16.0	44,454	20.7	317,016	25.8	23,124	4.7
Total maize nuclear	467,646	100.0	214,782	100.0	1,227,191	100.0	493,590	100.0

[†]The total number of sequences in each category expressed as a percentage of the total of maize nuclear (repeat + non-repeat) sequences. A sequence was defined as having a significant match to a sequence in one of the databases if identity was greater than 95% over a length of at least 40 bp with an expect value less than or equal to 10^{-10} using BLAT.

[‡]Maize Assembled Genome Island Version 4.0 (MAGIv4.0) contigs and singletons database is a partial maize genome assembly of B73 genomic sequences (Kalyanaraman et al. 2007) (<http://magi.plantgenomics.iastate.edu/>).

[§]MAGIv4.0 contigs database differs from the MAGIv4.0 contigs and singletons database in that it only contains consensus sequences derived from two or more overlapping reads.

[¶]MAGIv4.0 contigs genes database consists of 61,428 pre-mature mRNA gene structures that were predicted via running FGENESH v2.6 on the 163,390 MAGIv4.0 contig sequences. Gene structures consist of predicted UTRs, exons, and introns. In addition, the predicted gene structures are bordered by 300 bases upstream and downstream of the predicted transcription initiation and termination sites, respectively (<http://magi.plantgenomics.iastate.edu/>).

[#]The DFCI maize gene index (ZMGI release 17.0) consists of 115,744 unique expressed maize transcript sequences (<http://compbio.dfci.harvard.edu>).

distinguish recently diverged paralogs. However, it is unlikely that *HpaII* fragment sequences, with an average length of 120 bases after *in silico* digestion and a higher single-read error rate than that of Sanger sequencing, will contain sufficient and accurate information to distinguish between highly similar paralogs in the maize genome. In addition, if recently duplicated genes have diverged within the range of previously reported maize nucleotide diversity levels ($\pi=1-5\%$) (Henry and Damerval, 1997; Tenaillon et al., 2001; Wright et al., 2005), it will be difficult, if not impossible, to reliably distinguish paralogs based on the best reference match, reciprocal best match, or a conservative maximum allelic diversity threshold. Finally, the MAGIv4.0 C&S reference database used for SNP calling in this study is a partial genome assembly, thus the true allelic copy for an *HpaII* fragment sequence may not even be present in this reference database.

A two-pronged strategy was developed to deal with some of these challenges. First, the redundant and overlapping non-repeat B73 (husk and root: 61.3 Mb) and Mo17 (root: 133.3 Mb) *HpaII* fragment sequences (Table 3.2) were assembled into multiple sequence alignments and a consensus sequence representing each alignment was derived. Assembly of these sequences resulted in the derivation of 339,730 (42.6 Mb) and 586,237 (70.7 Mb) non-redundant *HpaII* consensus sequences from B73 and Mo17, respectively (Table 3.4). In addition to providing a longer assembled sequence to help accurately align *HpaII* fragments to allelic B73 reference sequences contained within the MAGIv4.0 C&S database (i.e., distinguish between highly similar paralogs), the assembly permitted a calculation of the per-base coverage depth, or the frequency with which any consensus base was observed in the raw data. Importantly, this metric can serve as a measure of confidence in the accuracy of consensus bases, as putative SNPs with a high coverage depth are more likely to be valid (Barbazuk et al.,

2007). In addition, the assembly of cognate *HpaII* fragment sequences reduced the computational requirements for the alignment and SNP calling process, as only unique sequences were used.

Table 3.4. Summary of the assembly process.

Coverage Depth [†]	B73 Husk and Root			Mo17 Root		
	No. [‡]	Mb [§]	% [¶]	No.	Mb	%
1	263,952	31.1	77.7	415,411	42.5	70.9
2	44,088	6.1	13.0	65,846	9.1	11.2
3	15,188	2.3	4.5	31,473	4.8	5.4
4	6,745	1.1	2.0	20,564	3.4	3.5
5+	9,757	2.0	2.9	52,943	10.9	9.0
Total	339,730	42.6	100.0	586,237	70.7	100.0

[†]Number of 454 sequences contained in each assembly.

[‡]Number of consensus sequences extracted from assemblies at each coverage depth.

[§]Number of consensus sequence bases at each coverage depth.

[¶]Percentage of the total number of consensus sequences at each coverage depth.

Second, we developed a computational approach to minimize the number of SNPs called from alignments of paralogous sequences, which is similar in objective to the paralog identification method used by the SNP calling software POLYBAYES (Marth et al., 1999) and to the “monoallelism” rules used by Barbazuk et al. (2007). Our approach assumes that it is possible to discover fixed differences among paralogs by comparing a reference sequence database or genome against itself, where almost all sequence differences observed in non-self paralogous alignments are non-allelic (Figure 3.1 A and B). Although some non-allelic differences may actually be polymorphisms at one or both of the loci, it is assumed that the majority of these identified differences are expected to be fixed differences that distinguish paralogs. Following this argument, a search of the MAGIv4.0 C&S database against itself was performed to identify all such single nucleotide differences that distinguish paralogs in the maize B73 genome. Putative non-allelic fixed differences that were identified from unique paralogous alignments were catalogued into a “paralog distinguishing list”

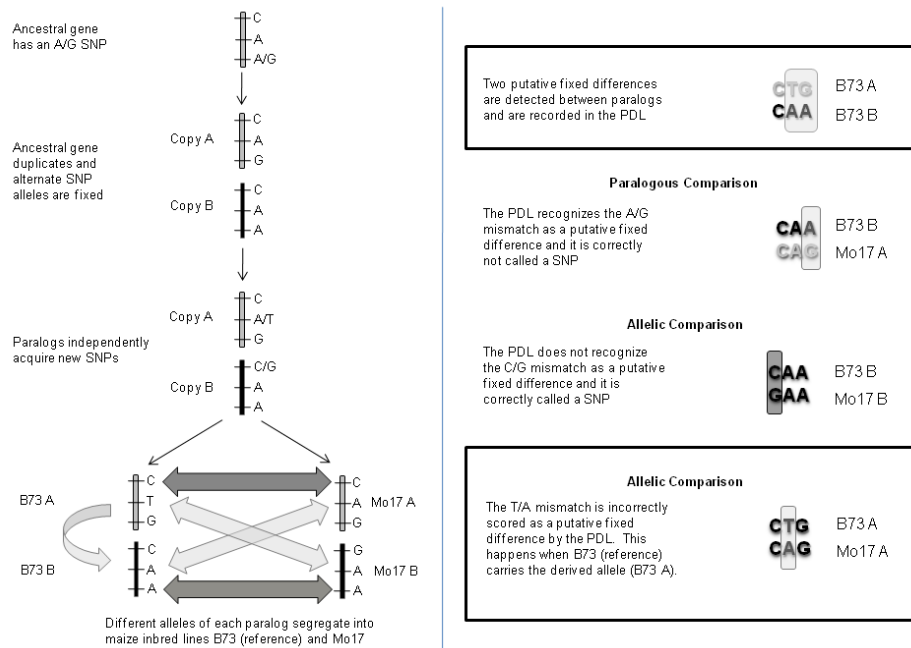


Figure 3.1. Illustration of a recent single gene duplication event that results in highly similar paralogs, and how the paralog distinguishing list (PDL) distinguishes alleles from paralogs when calling SNPs. (A) The PDL method is based on the assumption that a pair of duplicated genes that are fixed in the extant maize population likely originated from a single duplication event, which in many cases was the ancient tetraploidization event. If the duplication event is sufficiently old, virtually all differences among paralogs are because of mutations that have occurred since the genome duplication event, and distinguishing paralogs is easy. However, if the duplication was recent and the ancestral gene was polymorphic, alternative alleles at the paralogous loci may become fixed in the population, and the number of fixed differences between the donor and derived loci may be similar to the average allelic pairwise difference observed in maize. It is these cases for which it is very difficult to distinguish alleles from paralogs based only on alignment scores. (B) An intra-reference alignment of B73 reference sequences discovers putative fixed differences (T/A and G/A) that differentiate paralogs (B73 A and B73 B), which are recorded as context sequences in the paralog distinguishing list (PDL). Next, *HpaII* consensus sequences of Mo17 are aligned to B73 reference sequences. Both the correct allelic (B73 B vs. Mo17 B) and erroneous paralogous (B73 B vs. Mo17 A) alignments detect a single nucleotide mismatch, and thus, cannot be distinguished from each other based solely on alignment scores. The context sequences of both single nucleotide mismatches (A/G and C/G) are searched against the PDL. The context sequence of the A/G mismatch matches a context sequence in the PDL; thus, the mismatch is correctly recognized as a putative fixed difference and not called a SNP. However, the context sequence of the C/G mismatch does not match any context sequence in the PDL and is therefore correctly called a SNP. When B73 carries a derived allele (B73A), the context sequence of the T/A mismatch in the allelic B73 A vs. Mo17 A comparison is also detected in the PDL. Thus, this true SNP is not called because it is incorrectly scored as a putative fixed difference, which ultimately leads to a reduction in SNP detection power.

(PDL) as “context sequences” (i.e., the 16 bp 5' and 16 bp 3' flanking the single nucleotide difference).

SNP Identification

With the implementation of the PDL, *HpaII* consensus sequences from Mo17 were aligned against the best reference match B73 sequence (MAGIv4.0 C&S; 675.2 Mb) and all single nucleotide differences were identified and extracted as context sequences (see “Materials and Methods”). If the context sequence of *any* of these single nucleotide differences (Mo17 *HpaII* vs. B73 MAGIv4.0 C&S) matched a context sequence contained within the PDL, it was treated as an indication of a paralogous alignment and *all* SNP calls from such alignments were suppressed. In this case, the next strongest alignment for the same *HpaII* consensus sequence was considered, continuing in this fashion until an alignment with no match to a PDL context sequence was found, or the rate of mismatches in the successive alignments exceeded a preset maximum of 5%. Essentially, the PDL selected which alignments to use for SNP calling but not which single nucleotide differences to call as SNPs. The same procedure was performed with B73 *HpaII* consensus sequences, which served as an internal control to estimate the rate of false SNP discovery with and without implementation of the PDL.

Use of the PDL proved to be highly effective at preventing false SNP calls because of paralogous alignments. The estimated false discovery rate (FDR) obtained by comparing the SNP call rate for B73 (control, all SNPs considered false) and Mo17 *HpaII* consensus sequences at various coverage depths and base quality values (*Q* values) thresholds is shown in Table 3.5. If SNP calls were made using the PDL and not restricted to a specific coverage depth or *Q* value threshold, 126,683 putative SNPs between Mo17 and B73 (1 SNP/248 bp) were discovered at an estimated 15.1% FDR.

If SNP calls were made using only the most parsimonious alignment (i.e., without PDL), 174,476 putative B73/Mo17 SNPs (1 SNP/199 bp) were called at a dramatically increased FDR of 46.8%. Overall, use of the PDL effectively provided a 3-fold reduction in the rate of false SNP discovery at every evaluated coverage depth and Q value threshold relative to rates determined without use of the PDL.

As shown in Table 3.5, we observed a polymorphism rate of 1 SNP every 216 bp (86,830 SNPs/18,794,000 bp) at an estimated 11% FDR (Coverage Depth: $\geq 1X$; Q -score: ≥ 35). If we restricted SNP calling to a coverage depth of $\geq 2X$ (Q -score: all), then we observed a polymorphism rate of 1 SNP every 204 bp at a false SNP discovery rate of 8.4%. The SNP discovery rate for Mo17 *HpaII* consensus sequences at only 1X coverage (i.e., singletons) and all Q -scores was 1 SNP every 290 bp (calculated from Table 3.5) at an estimated 19.7% FDR, which suggests that at higher coverage depths and with higher quality sequence data more SNPs/kb were captured (i.e., higher SNP detection power). Although the FDR was reduced nearly 2-fold (15.1 to 8.4%) when using the PDL and additionally restricting SNP calls to a coverage depth of $\geq 2X$, the FDR remained relatively unchanged at progressively higher coverage depth thresholds. This suggests that deeper sequencing would provide limited improvement in the calling accuracy of SNPs already at a coverage depth of 2X or higher, but this might not have been the case if the sequenced maize lines were highly heterozygous. The ability to reduce the number of false positive SNPs by restricting SNP calls to higher cover depths was also a key finding by Barbazuk et al. (2007), the first study that used pyrosequencing to identify SNPs within expressed maize genes. Additionally, it seems that Q values calculated by the 454 base calling software (single reads) or CAP3 program (multiple sequence alignments) are of

Table 3.5. Summary of putative SNPs and call rates at various coverage depths and quality value thresholds with and without implementation of the paralog distinguishing list (PDL).

		With PDL					Without PDL				
CD [†]	Q [*]	B73		Mo17		FDR [¶]	B73		Mo17		FDR
		SNPs	Rate [§]	SNPs	Rate		SNPs	Rate	SNPs	Rate	
≥1X	All [#]	11,904	0.61	126,683	4.03	15.1%	50,936	2.35	174,476	5.02	46.8%
	≥20	10,701	0.58	119,294	4.02	14.4%	47,343	2.31	164,904	5.04	45.8%
	≥30	8,955	0.55	106,475	4.12	13.3%	39,910	2.23	147,335	5.16	43.2%
	≥35	5,703	0.51	86,830	4.62	11.0%	23,149	1.92	119,465	5.74	33.4%
	≥40	2,352	0.43	62,966	4.83	8.9%	10,378	1.78	85,547	5.92	30.1%
	≥50	1,609	0.37	57,205	4.93	7.5%	6,832	1.46	77,688	6.03	24.2%
	≥60	879	0.32	45,610	4.88	6.6%	3,724	1.26	61,991	5.97	21.1%
	≥70	634	0.30	39,787	4.88	6.1%	2,651	1.17	54,279	5.99	19.5%
≥2X	All	2,072	0.41	61,584	4.91	8.4%	9,048	1.66	83,547	6.00	27.7%
	≥20	2,057	0.41	61,527	4.91	8.4%	9,017	1.65	83,475	6.00	27.5%
	≥30	2,031	0.40	61,300	4.91	8.1%	8,910	1.64	83,173	6.00	27.3%
	≥40	1,953	0.40	60,573	4.91	8.1%	8,529	1.61	82,169	6.00	26.8%
	≥50	1,609	0.37	57,205	4.93	7.5%	6,832	1.46	77,688	6.03	24.2%
	≥60	879	0.32	45,610	4.88	6.6%	3,724	1.26	61,991	5.97	21.1%
	≥70	634	0.30	39,787	4.88	6.1%	2,651	1.17	54,279	5.99	19.5%
≥3X	All	702	0.33	37,980	4.88	6.8%	3,127	1.37	51,769	5.98	22.9%
	≥20	699	0.33	37,975	4.88	6.8%	3,124	1.37	51,763	5.98	22.9%
	≥30	697	0.33	37,966	4.88	6.8%	3,114	1.37	51,751	5.98	22.9%
	≥40	689	0.32	37,912	4.88	6.6%	3,088	1.36	51,681	5.98	22.7%
	≥50	679	0.32	37,833	4.88	6.6%	3,047	1.35	51,572	5.98	22.6%
	≥60	649	0.32	37,448	4.87	6.6%	2,899	1.32	51,044	5.97	22.1%
	≥70	529	0.30	35,417	4.87	6.2%	2,299	1.21	48,339	5.97	20.3%
≥4X	All	322	0.31	24,454	4.81	6.4%	1,452	1.31	33,403	5.90	22.2%
	≥20	319	0.31	24,454	4.81	6.4%	1,449	1.30	33,402	5.90	22.0%
	≥30	318	0.31	24,454	4.81	6.4%	1,445	1.30	33,402	5.90	22.0%
	≥40	317	0.30	24,451	4.81	6.2%	1,443	1.30	33,399	5.90	22.0%
	≥50	316	0.30	24,443	4.81	6.2%	1,437	1.30	33,391	5.90	22.0%
	≥60	313	0.30	24,430	4.81	6.2%	1,426	1.29	33,368	5.90	21.9%
	≥70	311	0.30	24,356	4.81	6.2%	1,405	1.28	33,272	5.90	21.7%

TABLE 3.5 (Continued)

[†]CD, coverage depth. The number of reads with an aligned base that supported the consensus base call.

[‡]*Q*, quality values. Quality values were computed using the 454 base-calling software (single reads) or the CAP3 assembly program (multiple sequence alignments).

[§]The number of SNPs called per kb of *Hpa*II consensus sequence (SNPs/kb).

[¶]The percent false discovery rate (FDR) at each coverage depth was calculated by dividing the B73 call rate by the Mo17 call rate and multiplying by 100.

[#]No filtering on *Q* values.

minimal value for eliminating false positive SNPs that result from sequencing errors when SNP calls are restricted to a coverage depth of 2X or higher.

SNP Validation

To independently cross-validate a subset of B73/Mo17 *HpaII* SNPs that were identified via 454 pyrosequencing, we extracted a collection of B73 and Mo17 amplicon sequences from the Panzea database (<http://www.panzea.org/>) (Zhao et al., 2006) that were generated with traditional Sanger sequencing chemistry. The extracted sequences were assembled and aligned according to unique Panzea locus identifiers, which permitted the identification of SNPs. It was assumed that all paired sequences were allelic and all true SNPs were identified (i.e., 0% FDR; 100% power). To estimate an FDR for *HpaII* SNPs, Panzea SNPs were mapped onto Mo17 *HpaII* consensus sequences, and vice versa. The mapping resulted in the identification of a subset of SNPs in each dataset that was derived from sequence common to both datasets (Table 3.6).

Table 3.6. Summary of B73/Mo17 454 SNP validation.

	With PDL	Without PDL
Panzea SNPs [†]	724	724
<i>HpaII</i> SNPs	523 [‡]	720 [§]
Shared SNPs [¶]	449	586
<i>HpaII</i> FDR [#]	14.1%	18.6%
<i>HpaII</i> Power ^{††}	62.0%	80.9%

[†]The number of identified Panzea SNPs that mapped to Mo17 *HpaII* consensus sequences.

[‡]The number of B73/Mo17 *HpaII* SNPs identified via 454 pyrosequencing that mapped to Panzea sequences. These B73/Mo17 *HpaII* SNPs that mapped are a subset of the 126,683 putative SNPs ($\geq 1X$ coverage depth; All *Q* values) that were called using the paralog distinguishing list (PDL).

[§]The number of B73/Mo17 *HpaII* SNPs identified via 454 pyrosequencing that mapped to Panzea sequences. These B73/Mo17 *HpaII* SNPs that mapped are a subset of the 174,476 putative SNPs ($\geq 1X$ coverage depth; All *Q* values) that were called without using the paralog distinguishing list (PDL).

[¶]SNPs that were identified in both the B73/Mo17 *HpaII* SNP and Panzea SNP datasets.

[#]We assumed that all SNPs called from the Panzea sequence dataset were true SNPs. The percent False Discovery Rate (FDR) was calculated as $[1-(449/523)*100]$ and $[1-(586/720)*100]$.

^{††}We assumed that all SNPs in the Panzea sequence dataset were identified. Power was calculated as $[(449/724)*100]$ and $[(586/724)*100]$.

With the constructed SNP validation dataset, we found that 85.9% (449/523) of the PDL-based *HpaII* SNPs were concordant with Panzea SNPs. This resulted in an estimated FDR of 14.1%, which strongly agreed with the 15.1% (no thresholds; with PDL) that was estimated using the B73/Mo17 call rate comparison (Table 3.5). However, only 62.0% of SNPs identified in Panzea were also identified in the dataset of PDL identified B73/Mo17 *HpaII* SNPs, whereas it was 80.9% without the PDL. This signifies a weakness of the MAGIv4.0 C&S-based PDL, as true SNPs were incorrectly considered non-allelic by the PDL.

DISCUSSION

Next generation DNA sequencing technologies have made high-throughput resequencing efficient and affordable. However, the use of these technologies in a read-to-reference based SNP discovery approach at the level of a whole-genome has not come to fruition for agronomically important plant species. The primary reason is that many of these plant species have large, complex genomes and as a result do not have an available, accurate or complete genome sequence. In addition, the short read lengths produced by these high-throughput sequencing technologies are limited in ability to differentiate the large numbers of paralogs that are common to the genome of many angiosperm species (Blanc and Wolfe, 2004). Maize was chosen as the test organism for this pilot study because of three qualities of its nuclear genome: it is ~2500 Mb in size; it consists of more than 75% highly repetitive DNA (Meyers et al., 2001; SanMiguel et al., 1996); and at least one-third of its estimated 59,000 genes are duplicated (Blanc and Wolfe, 2004; Messing et al., 2004). Here, we tested a gene-enrichment sequencing approach that is applicable to virtually any plant species and a

computational pipeline that enables the efficient and accurate discovery of a large number of SNPs using an incomplete and low-coverage reference sequence.

We modified the previously described HMPR technique (Emberton et al., 2005) to enable shotgun sequencing of 100-600 bp *HpaII* fragments in a manner that fully used the read length (potential of 200-300 bases) ability of the 454 GS FLX instrument. Of the two tissue types that were tested as sources of genomic DNA, seedling roots have a greater potential to enable the rapid construction of gene-enriched, modified HMPR libraries that have low levels of repeats and organellar DNA contamination. However, improved seed sterilization procedures and/or sterile, antibiotic-treated growing conditions are necessary to prevent the proliferation of bacterial symbionts in seedling roots, and the cytosine methylation pattern of genes and repeats in seedling root tissue needs to be more fully investigated. Since performing this experiment, we have identified unfertilized, immature ear shoots as an excellent tissue for isolating total maize genomic DNA. B73 and Mo17 immature ear *HpaII* libraries constructed with modified HMPR technology were highly enriched (4–5-fold) for genic sequences, while extremely depleted in repeat, organellar, and bacterial sequences (total: <10%) (M. Gore, R. Elshire, and E. Buckler, unpublished data).

Although our modified HMPR technique facilitated high throughput gene-enrichment sequencing of a large, complex plant genome, in general, the yield per run of modified HMPR libraries on the 454 GS FLX was lower than the expected 100 Mb. If the DNA copy per bead ratio is carefully optimized for modified HMPR libraries, it should be possible to routinely obtain 100 Mb of sequence data. In addition, the low sequencing yield may be because of less than optimal lengths (3-10 kb) of *HpaII* concatemers. If so, a 6 bp MCS restriction enzyme (Fellers, 2008) may help to produce much larger concatemers that are better suited for the downstream 454 sample

preparation, which is optimized for undigested total genomic DNA. Also, assembly of the larger restriction fragment sizes would produce larger consensus sequences for more accurate mapping. Alternatively, with the increased average read length (400 bases) and paired-end read capability of the new GS FLX Titanium (<http://www.454.com>), it might be more efficient and as comprehensive to directly sequence restriction fragments instead of concatemers.

We identified 126,683 putative B73/Mo17 SNPs, primarily in genic regions of the maize genome, using a computational pipeline for short read lengths that is applicable to any plant species with at least a large collection of genome survey sequences. A computational approach was developed to distinguish between allelic and paralogous *HpaII* consensus-MAGIv4.0 C&S reference alignments by searching identified putative single nucleotide differences against a Paralog Distinguishing List of putative fixed differences that distinguish paralogs from each other. The false SNP discovery rate with implementation of the PDL was estimated by two different approaches, and both were found to be at an acceptable level and highly concordant (15.1 vs. 14.1%). Detection of SNPs using the PDL was 3-fold more effective in controlling the FDR than a most parsimonious alignment strategy, and the FDR could be further reduced by filtering SNPs based on coverage depth and/or *Q* value thresholds (Table 3.5). The most likely sources of false positive SNPs are cloning artifacts (i.e., base substitution errors) contained within MAGIv4.0 C&S sequences (Fu et al. 2004) and paralogous alignments not identified by the PDL. Although very stringent parameters were used to assemble redundant, overlapping *HpaII* fragment sequences, it is possible that collapsed paralogs also contributed to the identification of false positive SNPs. The number of false positive SNPs that result from the FLX system are expected to be low (presumably less frequent at coverage depths of 2X and higher), as other studies have shown the GS FLX single-read error rate to be ~0.5%

(Droege and Hill, 2008) and substantially lower at higher coverage depths (Lynch et al., 2008; Smith et al., 2008). In addition, the rate of paralog collapse in the MAGI assemblies was estimated to be ~1% (Emrich et al., 2007); therefore, their contribution to the calling of false positive SNPs and inaccuracies in the PDL should be very minimal.

The difference in FDR estimates between SNPs called with and without the PDL method is much less striking for the Panzea validation dataset (Table 3.6) than that observed for the B73/Mo17 call rate comparison (Table 3.5). This is most likely because Panzea sequences resulted from the preferential sequencing of putatively single-locus PCR products, as PCR reactions that appeared to amplify multiple loci were discarded prior to sequencing (E. Buckler, unpublished). Essentially, the amplicon-Sanger sequencing strategy acted as a PDL. Thus, the Panzea dataset is poorly suited to assess the ability of the PDL to detect paralogous alignments, because the Panzea database was constructed with a bias against paralogous sequences. All amplicon-Sanger sequencing strategies will have this same bias; therefore, the best external validation of the PDL is to sequence modified HMPR libraries of Mo17 on a different next-generation sequencing platform (e.g., Illumina sequencing). Currently, the B73 (internal control)/Mo17 call rate comparison is the best available method to estimate the ability of the PDL to reduce the number of false positive SNP calls from paralogous alignments (Table 3.5). Nevertheless, minor improvements in the FDR are still observed when the PDL is used on the Panzea dataset (Table 3.6).

Transcriptome sequencing is useful when the aim is enrichment of tissue and developmental-stage specific genes; however, for high coverage of the gene space it is not very cost effective. Essentially, numerous cDNA libraries capturing multiple developmental stages and environmental stresses are needed to even approach high coverage of the gene space. Therefore, we sequenced modified HMPR genomic

libraries because it is expected to result in a more comprehensive sampling of genes than that of transcriptome sequencing (Emberton et al., 2005; Palmer et al., 2003), and it is also expected to provide access to the nucleotide diversity in introns, regulatory regions, and non-expressed genes. We used the Lander-Waterman model (Lander and Waterman, 1988) and the rate of contig formation as described in Whitelaw et al. (2003) to estimate the effective gene space size sampled by the modified HMPR method, which was 136.4 Mb (~27% of the ~500 Mb maize gene space; Palmer et al., 2003) for the Mo17 root library. This estimate of the effective gene space size might be slightly overestimated due to the very stringent CAP3 assembly parameters that were used. Given that 70.7 Mb of *HpaII* consensus sequence data exists for Mo17 (Table 3.4), it is estimated that the library was sequenced to only 0.52X coverage. If we were to sequence the Mo17 root library to 1X coverage, then the maximum number of putative SNPs called with the PDL would be ~200,000 at a rate of 4.03 SNPs/kb. If several million SNPs are to be discovered, we will need to sequence additional maize inbred lines, possibly construct other modified HMPR libraries using different 4 bp cutter MCS restriction enzymes, and/or use the draft maize genome sequence to call SNPs.

The PDL is only as high-quality as the completeness and accuracy of the reference sequence used to construct it, but despite the shortcomings of the MAGI assemblies (e.g., 1% collapsed paralogs, cloning artifacts, and partial genome assembly), a significant reduction (3-fold) in the number of false positive SNPs that resulted from paralogous alignments was still observed (Table 3.5). Moreover, these issues will be mostly resolved when the draft maize B73 genome sequence is available for constructing a PDL and calling SNPs.

A more important limitation of the PDL, however, is that it reduced the power to detect true SNPs. Based on the observed SNP call rate (4.91 SNPs/kb; 1 SNP/204

bp) with the PDL at a coverage depth of $\geq 2X$, we are under-estimating the expected SNP call rate (1 SNP/153 bp based on 1,095 genes) between any randomly chosen diverse, temperate maize inbred lines by $\sim 25\%$ (Yamasaki et al., 2005). If SNPs were called without the PDL at a coverage depth of $\geq 2X$, the observed (6.00 SNPs/kb; 1 SNP/167 bp) and expected (1 SNP/153 bp) SNP call rates are nearly identical. As shown in Table 3.6, based on the comparison of B73/Mo17 *HpaII* SNPs (no threshold) with the Panzea SNP dataset, there was an 18.9% loss in SNP detection power with implementation of the PDL. The reduction in power is attributed to true SNPs being incorrectly considered non-allelic by the PDL. We hypothesize that these true SNPs could not be distinguished from actual fixed differences among paralogs on the basis of the intra-reference sequence comparison alone, which would occur if the reference line (B73) used to construct the PDL carries a derived allele (Figure 1 A and B). This is a systematic bias that may affect both population genetics and association studies when the reference line alone carries an allele of interest. This problem is most severe when a single line is compared to the reference, but the expected rate of false negatives because of this effect decreases to $1/(n + 1)$ when n lines are compared to the reference. Further reduction may be possible if multiple non-reference lines are also compared to each other.

Although the results obtained in this pilot study are very encouraging, there are several drawbacks to this approach that should be considered. First, the method of gene enrichment used here restricts SNP discovery to sites near *HpaII* restriction sites in unmethylated regions, which can be remedied by constructing additional modified HMPR libraries with different 4 bp cutter MCS restriction enzymes. We do not presume that *all* nucleotide variation in methylated regions of the maize genome is phenotypically irrelevant, so different methods are needed to discover SNPs from these regions. Additionally, genome wide methylation patterns and locus specific

methylation levels may vary across genetic backgrounds, tissue types, developmental stages, and even environmental conditions (Cervera et al., 2002; Finnegan et al., 2000; Lister et al., 2008; Rabinowicz et al., 1999; Vaughn et al., 2007). Thus, performing this technique across a panel of inbred lines may not result in representation of all lines at all loci. For marker discovery, this line-specific or locus-specific censoring effect may not be important overall, but population genetic studies may be adversely affected by non-random missing data.

Regardless of these limitations, a considerable number of SNPs were discovered at an acceptably low FDR for the purpose of constructing high density multiplexed genotyping products, but sequencing of additional maize inbred lines is needed to construct a SNP dataset with low ascertainment bias that is appropriate for phylogenetics or population genetics studies. However, the SNPs identified in this study are immediately applicable for fine mapping of complex traits in the Intermated B73 x Mo17 (IBM) population, which is a widely used community resource for QTL mapping studies in maize (Lee et al., 2002). Most importantly, we estimate the cost of SNP discovery in this study at \$0.38/SNP yet note that several aspects of the molecular methods used here can be optimized for much higher sequencing yield and broader genome coverage. Such optimization, combined with further advances in high throughput sequencing yield, longer read lengths, lower error rates, and cheaper run costs, can further reduce the cost of SNP discovery in diverse maize such that several million gene-enriched SNPs needed for comprehensive association studies is an immediate economic possibility.

REFERENCES

- Altschul, S.F., T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D.J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- Antequera, F., and A.P. Bird. 1988. Unmethylated CpG islands associated with genes in higher plant DNA. *Embo J.* 7:2295-2299.
- Barbazuk, W.B., S.J. Emrich, H.D. Chen, L. Li, and P.S. Schnable. 2007. SNP discovery via 454 transcriptome sequencing. *Plant J.* 51:910-918.
- Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics* 5:433-438.
- Bennetzen, J.L., K. Schrick, P.S. Springer, W.E. Brown, and P. SanMiguel. 1994. Active maize genes are unmodified and flanked by diverse classes of modified, highly repetitive DNA. *Genome* 37:565-576.
- Blanc, G., and K.H. Wolfe. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Plant Cell* 16:1667-1678.
- Buckler, E.S., J.M. Thornsberry, and S. Kresovich. 2001. Molecular diversity, structure and domestication of grasses. *Genet. Res.* 77:213-218.
- Cervera, M.T., L. Ruiz-Garcia, and J.M. Martinez-Zapater. 2002. Analysis of DNA methylation in *Arabidopsis thaliana* based on methylation-sensitive AFLP markers. *Mol. Genet. Genomics* 268:543-552.
- Droege, M., and B. Hill. 2008. The Genome Sequencer FLX(TM) System--Longer reads, more applications, straight forward bioinformatics and more complete data sets. *J. Biotechnol.* 136:3-10.
- Emberton, J., J. Ma, Y. Yuan, P. SanMiguel, and J.L. Bennetzen. 2005. Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Res.* 15:1441-1446.
- Emrich, S.J., L. Li, T.-J. Wen, M.D. Yandea-Nelson, Y. Fu, L. Guo, H.-H. Chou, S. Aluru, D.A. Ashlock, and P.S. Schnable. 2007. Nearly Identical Paralog: Implications for Maize (*Zea mays* L.) Genome Evolution. *Genetics* 175:429-439.
- Fellers, J.P. 2008. Genome Filtering Using Methylation- Sensitive Restriction Enzymes with Six Base Pair Recognition Sites. *The Plant Genome* 1:146-152.

- Finnegan, E.J., W.J. Peacock, and E.S. Dennis. 2000. DNA methylation, a key regulator of plant development and other processes. *Curr. Opin. Genet. Dev.* 10:217-223.
- Fu, H., Z. Zheng, and H.K. Dooner. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl. Acad. Sci.* 99:1082-1087.
- Fu, H., W. Park, X. Yan, Z. Zheng, B. Shen, and H.K. Dooner. 2001. The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. *Proc. Natl. Acad. Sci.* 98:8903-8908.
- Gaut, B.S., and J.F. Doebley. 1997. DNA sequence evidence for the segmental allotetraploid origin of maize. *Proc. Natl. Acad. Sci.* 94:6809-6814.
- Gore, M., P. Bradbury, R. Hogers, M. Kirst, E. Verstege, J. van Oeveren, J. Peleman, E. Buckler, and M. van Eijk. 2007. Evaluation of Target Preparation Methods for Single-Feature Polymorphism Detection in Large Complex Plant Genomes. *Crop Sci.* 47:S-135-148.
- Hake, S., and V. Walbot. 1980. The Genome Of *Zea-Mays*, Its Organization And Homology To Related Grasses. *Chromosoma* 79:251-270.
- Henry, A.M., and C. Damerval. 1997. High rates of polymorphism and recombination at the Opaque-2 locus in cultivated maize. *Mol. Gen. Genet.* 256:147-157.
- Huang, X., and A. Madan. 1999. CAP3: A DNA Sequence Assembly Program. *Genome Res.* 9:868-877.
- Kalyanaraman, A., S.J. Emrich, P.S. Schnable, and S. Aluru. 2007. Assembling genomes on large-scale parallel computers. *J. Parallel Distr. Com.* 67:1240-1255.
- Kent, W.J. 2002. BLAT---The BLAST-Like Alignment Tool. *Genome Res.* 12:656-664.
- Lai, J., J. Ma, Z. Swigoňová, W. Ramakrishna, E. Linton, V. Llaca, B. Tanyolac, Y.-J. Park, O.Y. Jeong, J.L. Bennetzen, and J. Messing. 2004. Gene Loss and Movement in the Maize Genome. *Genome Res.* 14:1924-1931.
- Lander, E.S. and M.S. Waterman. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231-239.

- Lee, M., N. Sharopova, W.D. Beavis, D. Grant, M. Katt, D. Blair, and A. Hallauer. 2002. Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. *Plant Mol. Biol.* 48:453-461.
- Li, W., S. Ruf, and R. Bock. 2006. Constancy of organellar genome copy numbers during leaf development and senescence in higher plants. *Mol. Genet. Genomics* 275:185-192.
- Lister, R., R.C. O'Malley, J. Tonti-Filippini, B.D. Gregory, C.C. Berry, A.H. Millar, and J.R. Ecker. 2008. Highly Integrated Single-Base Resolution Maps of the Epigenome in *Arabidopsis*. *Cell* 133:523-536.
- Lynch, M., W. Sung, K. Morris, N. Coffey, C.R. Landry, E.B. Dopman, W.J. Dickinson, K. Okamoto, S. Kulkarni, D.L. Hartl, and W.K. Thomas. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc. Natl. Acad. Sci.* 105:9272-9277.
- Mardis, E.R. 2008. The impact of next-generation sequencing technology on genetics. *Trends Genet.* 24:133-141.
- Margulies, M., M. Egholm, W.E. Altman, S. Attiya, J.S. Bader, L.A. Bemben, J. Berka, M.S. Braverman, Y.J. Chen, Z. Chen, S.B. Dewell, L. Du, J.M. Fierro, X.V. Gomes, B.C. Godwin, W. He, S. Helgesen, C.H. Ho, G.P. Irzyk, S.C. Jando, M.L. Alenquer, T.P. Jarvie, K.B. Jirage, J.B. Kim, J.R. Knight, J.R. Lanza, J.H. Leamon, S.M. Lefkowitz, M. Lei, J. Li, K.L. Lohman, H. Lu, V.B. Makhijani, K.E. McDade, M.P. McKenna, E.W. Myers, E. Nickerson, J.R. Nobile, R. Plant, B.P. Puc, M.T. Ronan, G.T. Roth, G.J. Sarkis, J.F. Simons, J.W. Simpson, M. Srinivasan, K.R. Tartaro, A. Tomasz, K.A. Vogt, G.A. Volkmer, S.H. Wang, Y. Wang, M.P. Weiner, P. Yu, R.F. Begley, and J.M. Rothberg. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature* 437:376-380.
- Marth, G.T., I. Korf, M.D. Yandell, R.T. Yeh, Z. Gu, H. Zakeri, N.O. Stitzel, L. Hillier, P.-Y. Kwok, and W.R. Gish. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* 23:452-456.
- Messing, J., A.K. Bharti, W.M. Karlowski, H. Gundlach, H.R. Kim, Y. Yu, F. Wei, G. Fuks, C.A. Soderlund, K.F.X. Mayer, and R.A. Wing. 2004. Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci.* 101:14349-14354.
- Meyers, B.C., S.V. Tingey, and M. Morgante. 2001. Abundance, Distribution, and Transcriptional Activity of Repetitive Elements in the Maize Genome. *Genome Res.* 11:1660-1676.

- Palmer, L.E., P.D. Rabinowicz, A.L. O'Shaughnessy, V.S. Balija, L.U. Nascimento, S. Dike, M. de la Bastide, R.A. Martienssen, and W.R. McCombie. 2003. Maize genome sequencing by methylation filtration. *Science* 302:2115-2117.
- Possingham, J.V. 1980. Plastid Replication and Development in the Life Cycle of Higher Plants. *Ann. Rev. Plant Physiol.* 31:113-129.
- Rabinowicz, P.D. 2003. Constructing gene-enriched plant genomic libraries using methylation filtration technology. *Methods Mol. Biol.* 236:21-36.
- Rabinowicz, P.D., L.E. Palmer, B.P. May, M.T. Hemann, S.W. Lowe, W.R. McCombie, and R.A. Martienssen. 2003. Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res.* 13:2658-2664.
- Rabinowicz, P.D., K. Schutz, N. Dedhia, C. Yordan, L.D. Parnell, L. Stein, W.R. McCombie, and R.A. Martienssen. 1999. Differential methylation of genes and retrotransposons facilitates shotgun sequencing of the maize genome. *Nat. Genet.* 23:305-308.
- Rabinowicz, P.D., R. Citek, M.A. Budiman, A. Nunberg, J.A. Bedell, N. Lakey, A.L. O'Shaughnessy, L.U. Nascimento, W.R. McCombie, and R.A. Martienssen. 2005. Differential methylation of genes and repeats in land plants. *Genome Res.* 15:1431-1440.
- Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci.* 98:11479-11484.
- SanMiguel, P., A. Tikhonov, Y.K. Jin, N. Motchoulskaia, D. Zakharov, A. Melake-Berhan, P.S. Springer, K.J. Edwards, M. Lee, Z. Avramova, and J.L. Bennetzen. 1996. Nested retrotransposons in the intergenic regions of the maize genome. *Science* 274:765-768.
- Shendure, J., G.J. Porreca, N.B. Reppas, X. Lin, J.P. McCutcheon, A.M. Rosenbaum, M.D. Wang, K. Zhang, R.D. Mitra, and G.M. Church. 2005. Accurate Multiplex Polony Sequencing of an Evolved Bacterial Genome. *Science* 309:1728-1732.

- Smith, D.R., A.R. Quinlan, H.E. Peckham, K. Makowsky, W. Tao, B. Woolf, L. Shen, W.F. Donahue, N. Tusneem, M.P. Stromberg, D.A. Stewart, L. Zhang, S.S. Ranade, J.B. Warner, C.C. Lee, B.E. Coleman, Z. Zhang, S.F. McLaughlin, J.A. Malek, J.M. Sorenson, A.P. Blanchard, J. Chapman, D. Hillman, F. Chen, D.S. Rokhsar, K.J. McKernan, T.W. Jeffries, G.T. Marth, and P.M. Richardson. 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Res.* 18:1638-1642.
- Smith, T.F., and M.S. Waterman. 1981. Identification of common molecular subsequences. *J. Mol. Biol.* 147:195-197.
- Swigoňová, Z., J. Lai, J. Ma, W. Ramakrishna, V. Llaca, J.L. Bennetzen, and J. Messing. 2004. Close Split of Sorghum and Maize Genome Progenitors. *Genome Res.* 14:1916-1923.
- Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci.* 98:9161-9166.
- Vaughn, M.W., Tanurd, Milo, Z. Lippman, H. Jiang, R. Carrasquillo, P.D. Rabinowicz, N. Dedhia, W.R. McCombie, N. Agier, A. Bulski, V. Colot, R.W. Doerge, and R.A. Martienssen. 2007. Epigenetic Natural Variation in *Arabidopsis thaliana*. *PLoS Biology* 5:e174.
- Wei, F., E. Coe, W. Nelson, A.K. Bharti, F. Engler, E. Butler, H. Kim, J.L. Goicoechea, M. Chen, S. Lee, G. Fuks, H. Sanchez-Villeda, S. Schroeder, Z. Fang, M. McMullen, G. Davis, J.E. Bowers, A.H. Paterson, M. Schaeffer, J. Gardiner, K. Cone, J. Messing, C. Soderlund, and R.A. Wing. 2007. Physical and Genetic Structure of the Maize Genome Reflects Its Complex Evolutionary History. *PLoS Genetics* 3:e123.
- Whitelaw, C.A., W.B. Barbazuk, G. Perte, A.P. Chan, F. Cheung, Y. Lee, L. Zheng, S. van Heeringen, S. Karamycheva, J.L. Bennetzen, P. SanMiguel, N. Lakey, J. Bedell, Y. Yuan, M.A. Budiman, A. Resnick, S. Van Aken, T. Utterback, S. Riedmuller, M. Williams, T. Feldblyum, K. Schubert, R. Beachy, C.M. Fraser, and J. Quackenbush. 2003. Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302:2118-2120.
- Wright, S.I., I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen, and B.S. Gaut. 2005. The effects of artificial selection on the maize genome. *Science* 308:1310-1314.

- Yamasaki, M., M.I. Tenaillon, I. Vroh Bi, S.G. Schroeder, H. Sanchez-Villeda, J.F. Doebley, B.S. Gaut, and M.D. McMullen. 2005. A Large-Scale Screen for Artificial Selection in Maize Identifies Candidate Agronomic Loci for Domestication and Crop Improvement. *Plant Cell* 17:2859-2872.
- Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeu, B.J. Nikolau, and P.S. Schnable. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. *Proc. Natl. Acad. Sci.* 99:6157-6162.
- Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2002. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. *Genome Res.* 12:1345-1349.
- Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2003. High-Cot sequence analysis of the maize genome. *Plant J.* 34:249-255.
- Zhao, W., P. Canaran, R. Jurkuta, T. Fulton, J. Glaubitz, E. Buckler, J. Doebley, B. Gaut, M. Goodman, J. Holland, S. Kresovich, M. McMullen, L. Stein, and D. Ware. 2006. Panzea: a database and resource for molecular and functional diversity in the maize genome. *Nucleic Acids Res.* 34:D752-757.
- Zhu, C., M. Gore, E.S. Buckler, and J. Yu. 2008. Status and Prospects of Association Mapping in Plants. *The Plant Genome* 1:5-20.

CHAPTER 4
A FIRST GENERATION HAPLOTYPE MAP OF THE WORLD'S MOST
DIVERSE CROP: MAIZE

ABSTRACT⁵

Maize (*Zea mays* L.) is the number one production crop in the world and possesses more genetic diversity than any other major crop species. This standing genetic variation is central to modern maize improvement, but is likely millions of years old, reflecting the historical evolutionary forces of recombination and selection as well as past demographic events. By using low-copy-enrichment and rapid sequencing-by-synthesis (SBS) approaches, we simultaneously discovered and genotyped nearly 3 million non-redundant polymorphisms in a diverse panel of 27 maize inbred lines. A substantial portion of the low-copy fraction in the maize genome was identified as being highly divergent or inserted relative to the reference genome sequence. We detected more than 20 massive regions (>800 Kb) of low diversity interspersed throughout the genome that are presumably the result of selection during maize evolution, which are substantially larger than known domestication loci. Genome-wide estimates of recombination rate based on sequence data (C) and a multi-population genetic linkage map (R) were strongly correlated with nucleotide diversity (π), hinting at a possible role of selection, along with genome structure, in patterning polymorphism throughout the genome. Additionally, while most of the genome shows no genetic differentiation (F_{ST}) between temperate and tropical germplasm, nearly one hundred regions are highly differentiated, likely containing loci key to geographic adaptation. Because these diverse lines are the founders of the largest set of public

⁵ M. A. Gore, J.-M. Chia, R. J. Elshire, J. Ross-Ibarra, Q. Sun, E. S. Ersoz, B. L. Hurwitz, J. A. Peiffer, G. S. Grills, D. H. Ware, E. S. Buckler. To be submitted to a journal with a high impact factor.

mapping populations for complex trait dissection—the maize Nested Association Mapping (NAM) population—this work also lays the foundation for truly Genome-Wide Association Studies (GWAS) in maize.

INTRODUCTION

Maize (*Zea mays* L.) is both a model genetics system and crop with high economic and societal value. Already an important source of food, fuel, feed, and fiber in the world, maize stands to be further improved through plant breeding practices that exploit the maize genome's genetic diversity. Maize has unparalleled genetic diversity for a model species, with the average nucleotide diversity of coding regions between any two maize lines ($\pi=1-1.4\%$) (Tenaillon et al., 2001; Wright et al., 2005) similar to the divergence between humans and chimpanzees (The Chimpanzee Sequence and Analysis Consortium, 2005). Maize also has tremendous phenotypic and geographic diversity—varieties of maize express a plethora of stable and plastic phenotypes and have adapted to distinct environments such as lowland tropics, hot deserts, high altitude mountains, and very short growing seasons.

Understanding the relationship between genetic and phenotypic variation is vital to manipulating and preserving maize diversity in this period of rapid climate change and increased global demands for water, land, energy, and food. Many important agronomic traits are genetically complex, however, and it has been difficult to connect phenotype to individual genes and alleles (Holland, 2007; Salvi and Tuberosa, 2005). Genome-Wide Association Studies (GWAS) using diverse maize germplasm offer the potential to rapidly resolve complex traits to a single gene or an individual polymorphism, but these studies require a high-density of genome-wide markers (Buckler et al., 2006; Yu and Buckler, 2006). To bolster maize as a model system for studying the genetic basis of complex traits, we have constructed a dataset

of 2.8 million polymorphism from a diverse panel of 27 inbred lines—founders of the maize Nested Association Mapping (NAM) population (Yu et al., 2008)—and have used this dataset to investigate the evolutionary forces shaping genetic diversity in maize.

MATERIALS AND METHODS

DNA isolation

Immature, unfertilized ears were harvested from multiple field-grown plants of 27 maize inbred lines: B73, B97, CML52-RIL, CML69, CML103, CML228, CML247, CML277, CML322, CML333, HP301, Il14H, Ki3, Ki11, Ky21, M37W, M162W, Mo17, Mo18W, MS71, NC350, NC358, Oh43, Oh7B, P39, Tx303, and Tzi8. The founder lines were chosen to maximize overall allelic richness, represent important public U.S. inbred lines, and permit the production of seeds in the U.S. summer (Yu et al., 2008). The CML52 line was not a pure genetic stock, as it was a recombinant inbred line (RIL) of mixed B73 and CML52 parentage. The CML52-RIL will be used to calibrate SNP scoring algorithms in the future. All collected ears were surface sterilized in a 10% (vol/vol) bleach solution (5.25% Sodium Hypochlorite) by gently rocking for 30 min, followed by 3X 10 min rinses with sterile water, and immediately frozen in liquid N₂ prior to storage at –80°C. Total genomic DNA was isolated from homogenized frozen ear tissue as previously described by Emberton et al. (2005).

Construction of Genomic DNA Libraries

Methylation-Filtration HpaII

In plant genomes, retrotransposons and other types of repeats contain a higher density of methylation in the form of 5-methylcytosine relative to genic sequences,

which is a property unique to plants (Rabinowicz et al., 2003; Rabinowicz et al., 2005). *HpaII*, a methylation-cytosine sensitive restriction enzyme, has been previously used to exploit this property for gene-enrichment sequencing via methylation-filtration (MF) of the maize genome (Emberton et al., 2005; Yuan et al., 2002). Complete digestions of 5 µg of high-molecular weight (HMW) total genomic DNA were performed in 200 µL volumes with 50 U of *HpaII* (New England Biolabs, Ipswich, MA) at 37°C for 16 h, followed by heat inactivation of the enzyme at 65°C for 20 min (Figure 4.1). Digestion reactions were performed in duplicate for each of the 27 lines, and duplicate reactions (i.e., identical genotype) were pooled following heat inactivation. All pooled samples were individually purified using the QIAquick PCR Purification Kit (QIAGEN, Valencia, CA), according to the manufacturer's protocol, and separated on a low melting 2% SeaPlaque agarose gel (Cambrex Bio Science Rockland, Inc., Rockland, ME). The gel was stained with a DNA visualization dye (SYBR Safe, Invitrogen Corp., Eugene, OR) and viewed on a blue fluorescent light box (Dark Reader, Clare Chemical Research, Denver, CO).

HpaII fragments ranging in size from 100–600 bp were excised from the agarose gel with a sterile scalpel. Gel-extracted fragments were purified and eluted with 40 µl of elution buffer using the QIAquick Gel Extraction Kit, according to the manufacturer's protocol. For each sample, the elutant of DNA was divided into four equal aliquots, with each 10 µl DNA aliquot separately digested with one of four different restriction enzymes. Complete digestions of the 10 µl DNA aliquots were performed in 35 µl volumes using 20 U of *AluI*, *HaeIII*, *MspI* or *RsaI* (New England Biolabs, Ipswich, MA) at 37°C for 16 h, followed by heat inactivation of the enzyme at 65°C for 20 min. The four separate digestion reactions were pooled and purified using the QIAquick PCR Purification Kit.

Whole-Genome Amplification *HpaII*

Differential methylation at *HpaII* recognition sites among inbred lines caused non-random missing sequence data from low-copy regions of the maize genome (data not shown). In an effort to sequence these differentially methylated, low-copy regions,

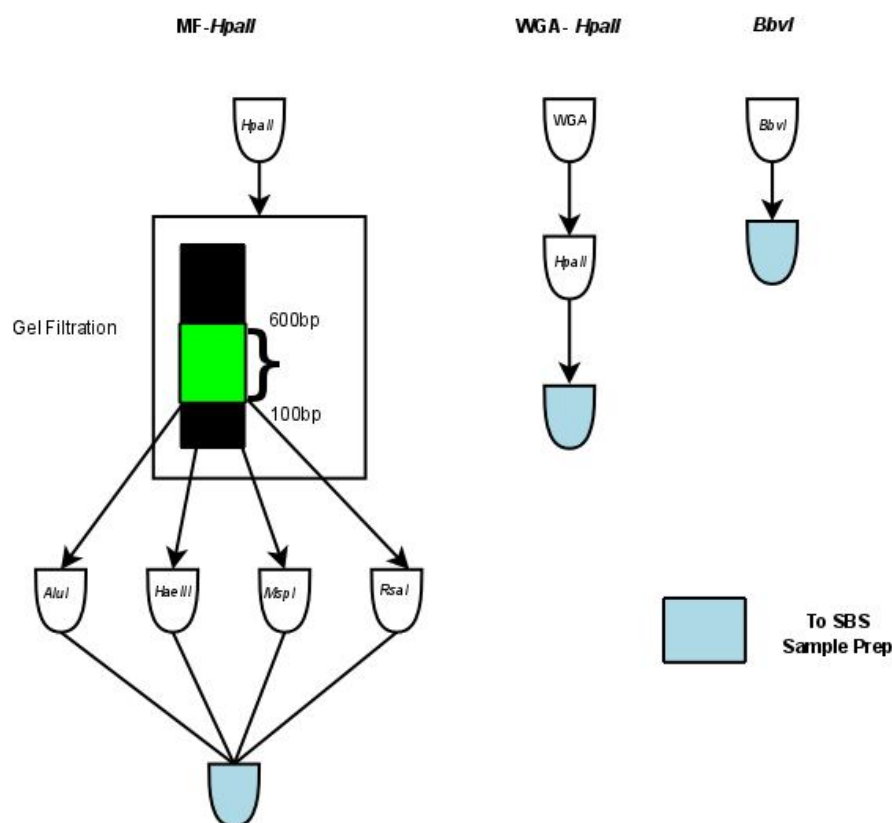


Figure 4.1. Flowchart of how methylation-filtration (MF) *HpaII*, whole-genome amplification (WGA) *HpaII*, and *BbvI* genomic libraries were constructed.

a whole-genome amplification (WGA) reaction was used to generate unmethylated genomic DNA. The GenomiPhi V2 DNA Amplification Kit (GE Healthcare, Piscataway, NJ) was used to amplify 1 μ L of 50 ng/ μ L of total genomic DNA per the manufacturer's instructions. This kit uses high fidelity Phi29 (ϕ 29) DNA polymerase, dNTPs, and random hexamers to replicate linear genomic DNA by multiple displacement amplification. A single WGA reaction was performed for each inbred

line and individually purified using the QIAquick PCR Purification Kit. Next, the unmethylated genomic DNA samples were completely digested with *HpaII* as described above, but they were subsequently purified using the QIAquick PCR Purification Kit without any prior size selection on an agarose gel.

Low-copy enriched BbvI

The methylation-insensitive, Type IIS restriction enzyme *BbvI* has a recognition site that occurs at higher frequency within low-copy relative to repetitive regions of the maize genome (data not shown). By constructing *BbvI* libraries, we were able to sequence additional low-copy regions not captured in *HpaII* libraries, and thus, increase sequencing coverage across the genome. Complete digestions of 5 µg of HMW total genomic DNA were performed in 200 µL volumes with 10 U of *BbvI* (New England Biolabs, Ipswich, MA) at 37°C for 16 h, followed by heat inactivation of the enzyme at 65°C for 20 min. The digestion reactions were purified using the QIAquick PCR Purification Kit.

Construction of Illumina SBS Libraries

Generation of single-end Illumina sequencing-by-synthesis (SBS) libraries was carried out according to manufacturer's instructions (Illumina, San Diego, CA), with minor modification. Briefly, a 3' adenosine overhang was added to polished DNA fragments using 1 µl of 1:5 diluted Klenow polymerase. Next, 1 µl of 1:10 diluted DNA oligonucleotide (oligo) adapters were ligated to DNA fragments of MF-*HpaII* libraries, whereas 1 µl of 1:5 diluted DNA oligo adapters were ligated to DNA fragments of WGA-*HpaII* and *BbvI* libraries. The adapter-ligated DNA was amplified using PCR primers 1.1 and 2.1 for 18 cycles, and purified using the QIAquick PCR Purification Kit. DNA fragments were quantified on a NanoDrop (NanoDrop Technologies, Wilmington, DE). In addition, an aliquot of the amplification reaction

was separated on a 1% agarose gel to calculate the mean size of amplicons. DNA was diluted to 10 nM using elution buffer (QIAGEN) supplemented with 0.1% Tween 20 and stored at -20°C .

Illumina SBS Sequencing

The Cluster Generation Kit was used to produce clusters on a Cluster Station per Illumina's instructions. Before samples were sequenced a second time, we recalibrated the concentration of the denatured DNA solution based on initial cluster numbers to optimize the number of raw clusters per tile. With the 36 Cycle Solexa Sequencing Kit (Illumina), samples were sequenced on an Illumina Genome Analyzer. Over several months of generating sequence data, multiple versions of the Illumina Genome Analyzer data pipeline (0.2.2.6; 0.3; and 1.0) and flow cells (1 and 2) were used. The first 36 bases of runs were extracted and processed. The data quality of each run was assessed, and runs without errors due to equipment malfunction or poor flow cell performance were used for further analysis. All data from runs which we deemed poor quality were discarded.

Mapping Reads to the Reference Genome Sequence

The ELAND software (Illumina, San Diego, CA), a short read alignment algorithm, was used to map SBS reads to the maize B73 reference genome sequence [Bacterial Artificial Chromosome (BAC) release 3a50]. ELAND trimmed the last 4 bases of reads, because the software only processes the first 32 bases. In addition, the 3' terminal bases of SBS reads tend to be relatively lower quality. The alignment parameters of ELAND allowed a maximum of 2 mismatches and no indels. Reads that mapped to 4 or less locations on the reference genome sequence were used to identify polymorphisms. The stringent ELAND alignment parameters [allowing up to 2

mismatches] resulted in a significant number of unmapped reads.

The Novoalign software (<http://www.novocraft.com/>), which has an alignment algorithm not limited to 2 mismatches, was used to align the unmapped reads. First, Novoalign finds possible alignment locations against an indexed reference sequence and proceeds to score the alignments using the Needleman-Wunsch algorithm with affine gap penalties. The reference genome sequence was indexed using a 14-mer seed length and a sliding window of 2 bp (-k 14, -s 2). Alignments were performed with a gap-opening penalty of 42, a gap-extension penalty of 4 and a maximum threshold of 85 (-g 42, -x 4, -t 85). Novoalign also used base quality scores to reduce misalignments. Only 3 bp indels could pass the filters given the length of reads and thresholds used for alignment. Also, reads that mapped to 4 or less positions were retained, but reads with 3 or more mismatches were discarded.

Identification of Putative Polymorphisms

ELAND and Novoalign alignment results were parsed and combined, and subsequently indexed on a per chromosome basis using a MySQL relational database and custom Perl scripts. Putative polymorphisms (SNPs and indels of 1-3 bp) were identified pairwise by collapsing mapped reads from each inbred line onto the reference genome sequence. In addition, read coverage depth of all genome bases (polymorphisms and non-polymorphisms) was determined.

Scoring of Polymorphisms

Putative polymorphisms were scored with a two-prong approach. First, a test of independence was used to identify a non-random association between alleles of a polymorphism among inbred lines. The null hypothesis is that the ratio of two alleles (read count of reference allele: read count of alternative allele) should be the same

among inbred lines in the absence of a segregating polymorphism. If the null hypothesis is rejected, there is a significant difference in the ratio of two alleles among inbred lines; thus, it is a segregating polymorphism (i.e., genotypic signal). For each detected putative SNP and indel, a 2 x 27 contingency table was built by counting the number of times each allele (i.e., read number) of an individual polymorphism was observed for an inbred line. All identified putative polymorphisms were treated as biallelic. If a putative polymorphism had three alleles (e.g., due to a sequencing error), the allele with the lowest frequency was discarded before populating the contingency table. In addition, columns pertaining to inbred lines without any reads were removed from the contingency table before implementing the test of independence. A Monte Carlo method was used for the empirical estimation of *P* values, with 1,000 random permutations and row and column totals fixed (Weir, 1996).

Second, a Fisher's exact test was used to determine if putative polymorphisms segregated in a manner that was consistent with expectations for a single-locus mapping polymorphism. Inbred maize lines have homozygous genotypes (e.g., AA or aa); therefore, the expectation is that individual inbred lines should be homozygous for either the reference or alternative allele (i.e., 100% of the reads should only report one allele). Segregating polymorphisms that map to multiple genomic positions will deviate from this homozygous expectation, which would most likely result from highly similar paralogous sequences. To determine the favored threshold for a given marker, we maximized the significance (*P* value) of the Fisher's exact test for each polymorphism by testing proportions that ranged from 0.01 to 0.96 at increments of 0.05. The significance of each proportion was compared to that obtained with the 0.5 threshold. Single-locus polymorphisms will have a maximum threshold value near 0.5, while paralogous polymorphisms will have a maximum threshold value near 0.25 (e.g., duplicated loci) or lower. Putative polymorphisms that had a maximum threshold

of 0.5 were scored as single-locus, while those that deviated from 0.5 were scored as paralogous.

Putative polymorphisms with a test of independence P value of ≤ 0.01 and scored on ≥ 13 lines were considered statistically significant, regardless of issues with paralogy. All of the identified statistically significant polymorphisms are useful for GWAS, but only single-locus markers are ideal for studies of population genetics.

Nucleotide Diversity and Genetic Differentiation

Only single-locus SNPs scored on ≥ 20 inbred lines were used for conducting population genetics analyses on 15,179 BACs. Pairwise nucleotide diversity (π) was estimated separately for each BAC per Nei (1987). Subsequently, a moving average of 9 BACs was applied to the estimates of pairwise nucleotide diversity, which simply served as a smoothing operation. All reported π estimates were per site values. To estimate the amount of genetic variation within and between tropical/subtropical (CML52-RIL, CML69, CML103, CML228, CML247, CML277, CML322, CML333, Ki3, Ki11, M37W, Mo18W, NC350, NC358, Tx303, and Tzi8) and temperate (B73, B97, HP301, Il14H, Ky21, M162W, Mo17, MS71, Oh43, Oh7B, and P39) subpopulations of maize, we calculated F_{ST} for each SNP per Weir and Cockerham (1984). Subsequently, F_{ST} values were averaged for each BAC.

Linkage Disequilibrium and Recombination

Linkage disequilibrium (LD) was estimated using squared allele-frequency correlations (r^2) for pairs of loci as previously described by Remington et al. (2001). Only sites with a frequency of at least 0.10 for the rarer allele were included. Mean r^2 was estimated for each BAC using pairs of SNPs at five different physical distances within a BAC: 1-10 bp, 11-100 bp, 101-1,000 bp, 1,001-10,000 bp, and 10,001-

100,000 bp. To calculate the population-recombination parameter C ($4N_e c$) for each BAC per Hudson (1987), $1/\text{mean } r^2$ was regressed on bp distance between sites. Similarly, a moving average of 9 BACs was applied to the C estimates. All reported C estimates were per site values. In addition, C estimates deemed as outliers were removed from the dataset. A physical estimate of recombination (R) was determined by calculating the ratio of genetic (cM) and physical (Mb) distances for every pair of contiguous markers on the NAM genetic linkage map (McMullen, in review). The genetic linkage map was constructed using the 25 linkage populations that comprise the maize NAM population (McMullen, in review). A 1 cM window was used to estimate R , and markers discordant between the physical and genetic maps were removed from the analysis.

Descriptive Statistics and Statistical Correlations

Descriptive statistics for π , LD, C , and F_{ST} were only reported for 12,357 BACs, which had at least 2,000 independent sites (polymorphic or non-polymorphic) aligned to SBS reads. Autocorrelations for estimates of π and C were calculated with the AUTOREG procedure in SAS (SAS Institute, Cary, NC). Best-fit, non-linear regression (Microsoft Excel, Microsoft Corp., Redmond, WA) was used to determine the fraction of total variance explained (R^2) by three models: C on π (power trendline), R on π (exponential trendline), and C on R (logarithmic trendline). Only a subset of 1,686 BACs (i.e., every ninth BAC) was used in the non-linear regression analysis, which helped to reduce the number of observations in the model and levels of autocorrelation. The C and π estimates used in the non-linear regression analysis were those calculated using the 9 BAC moving average described in the previous section.

Sequence Features and Stepwise Regression

BAC sequences (release 3a50) used for the sequence feature analysis were downloaded from the maize genome sequence database (<http://www.maizesequence.org>). CpG content and GC content were determined for unmasked BAC sequences. Repeat content was determined by counting the number of repeat masked bases for each masked BAC sequence. The number of genes per BAC was calculated using a list of protein-coding genes with corresponding BAC positions (<http://www.maizesequence.org>). The distance of each BAC from its respective centromere and telomere was calculated according to annotated physical positions of centromeres and telomeres (<http://www.maizegdb.org>). These data were joined with the subset of 1,686 BACs used in the non-linear regression analysis.

Multiple linear regression using a stepwise procedure was performed with the GLMSELECT procedure in SAS (SAS Institute, Cary, NC). The SELECT=ADJR SQ option was used, with maximum steps equal to the number of independent factors (6 or 7) in the model. R , C , or π was the dependent variable, while the six sequence features and R (only for C and π) were independent variables in a multiple linear regression model. The estimates of C , R and π were identical to those used in the non-linear regression analysis, thus the results from the multiple linear regression-stepwise procedure were based on the same subset of 1,686 BACs.

RESULTS AND DISCUSSION

Construction of a dense SNP map

To access the non-repetitive, recombinationally active regions of the maize genome, we constructed three types of restriction enzyme-anchored genomic libraries for each diverse line (see “MATERIALS AND METHODS”). Methylation-filtration (MF) *Hpa*II libraries were constructed to allow the preferential sequencing of regions

that flanked unmethylated *HpaII* recognition sites, which are more abundant in genic and low-copy-number regions of the maize genome (Emberton et al., 2005; Yuan et al., 2002). Substantial genotypic-specific methylation differences were observed at allelic *HpaII* sites across the 27 diverse lines, which caused non-random missing sequence data (Ersoz, Chia, Ware, and Buckler, unpublished data). To eliminate this bias, we also constructed *HpaII* libraries using genomic DNA that was unmethylated via whole-genome amplification (WGA) at the expense of an increase in highly repetitive sequences. In addition, libraries enriched for low-copy-number sequences were constructed using *BbvI*, a type IIS restriction enzyme that complemented *HpaII* without sensitivity to cytosine methylation.

We used Illumina sequencing-by-synthesis (SBS) technology (formerly known as Solexa sequencing) (Bennett, 2004) to resequence the “low-copy-number space” of the 27 lines. More than 1 billion SBS reads (>32 Gb) were generated via multiple sequencing runs on 81 genomic libraries. As expected, the majority of SBS reads from MF-*HpaII* and *BbvI* libraries mapped to unique positions on the unassembled, maize B73 reference genome sequence, while fewer low-copy-number SBS reads were observed in WGA-*HpaII* libraries (Figure 4.2). We achieved nearly 40-fold read

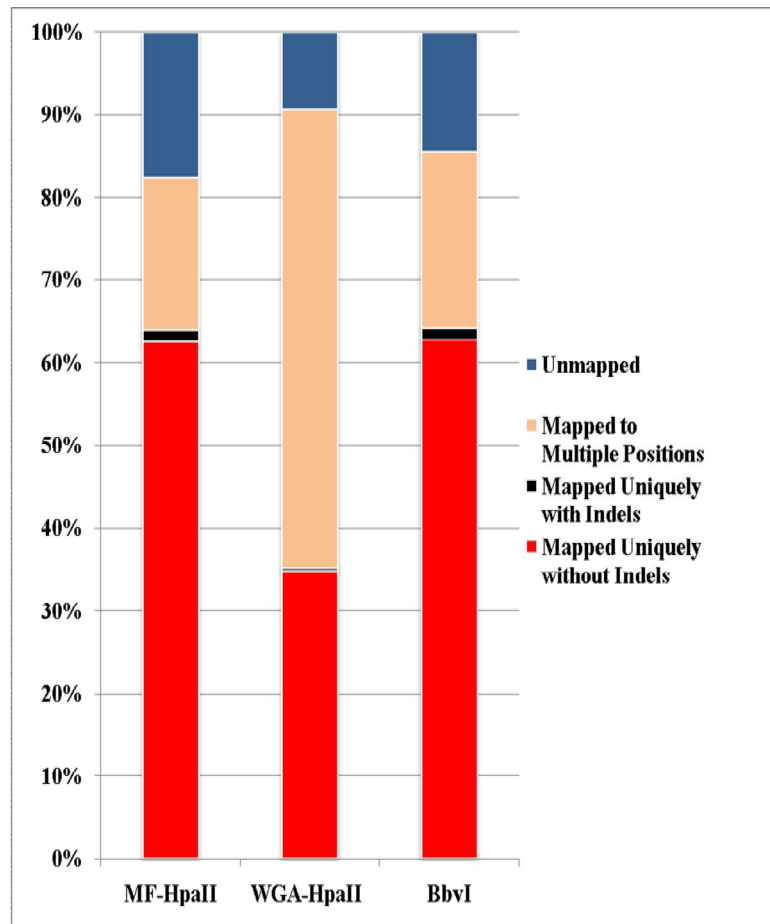


Figure 4.2. Percentage (%) of low-copy, high-copy, and unmapped SBS reads for all 27 diverse lines by library type. There were many instances of overlapping Bacterial Artificial Chromosome (BAC) sequences. Therefore, SBS reads that mapped to ≤ 4 positions on the unassembled, maize B73 reference genome sequence were considered unique, while those mapping to multiple positions (≥ 5) were deemed high-copy. SBS reads that could not be anchored to the reference genome sequence were considered unmapped. The observed proportions were expected for the MF-*HpaII* and WGA-*HpaII* libraries, but the *BbvI* library contained more than the expected number of uniquely mapping reads.

coverage depth of a low-copy-number fraction of the maize genome calculated to be 100 Mb in size, but in total about 45% of the nucleotide bases in the maize genome were covered by SBS reads—albeit at mostly low-coverage (Figure 4.3).

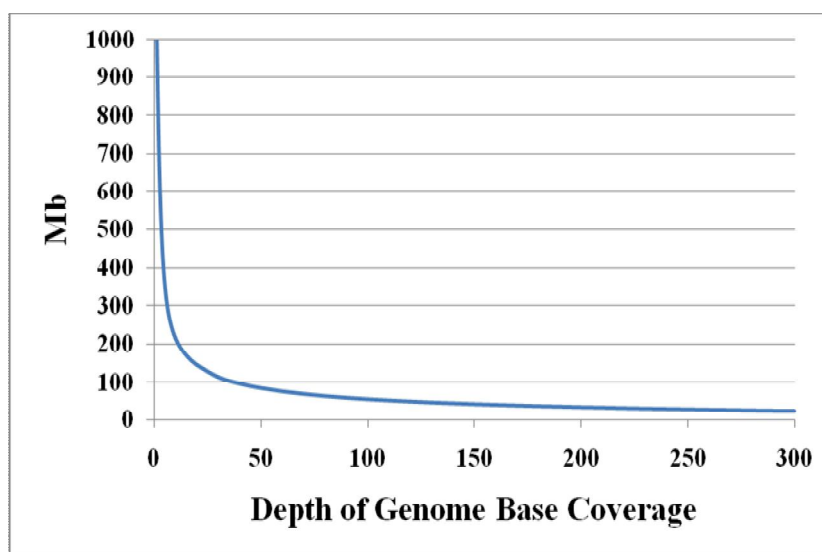


Figure 4.3. Number of Megabases (Mb) in the maize genome that was sequenced versus depth of genome base coverage. Coverage depth (i.e., oversampling of the genome) was calculated as the number of reads that aligned to a nucleotide base in the maize genome. Based on anchored BACs, the size of the B73 genome was estimated to be ~2400 Mb.

We used a custom polymorphism discovery pipeline to identify putative single-nucleotide polymorphisms (SNPs) and indels between aligned SBS reads and the reference genome sequence. All detected variants were scored with a test of independence to identify those most likely to be true polymorphisms (see “MATERIALS AND METHODS”). A subset of 2,867,766 non-redundant SNPs and indels was successfully genotyped based on our set criteria, with these polymorphisms scored on ≥ 13 lines and mapping to positions that were not highly repetitive. In total, 63,523,497 individual genotypes were determined for these SNPs and indels, with an average of 22 lines scored per polymorphism. There were no missing data for 25.1% of these polymorphisms, and 70.1% of them were scored on ≥ 20 lines. In addition, nearly 5% of the maize genome in ≥ 13 lines was covered by these polymorphisms with an average and median BAC inter-marker distance of 811 and 23 bp, respectively (Figure 4.4).

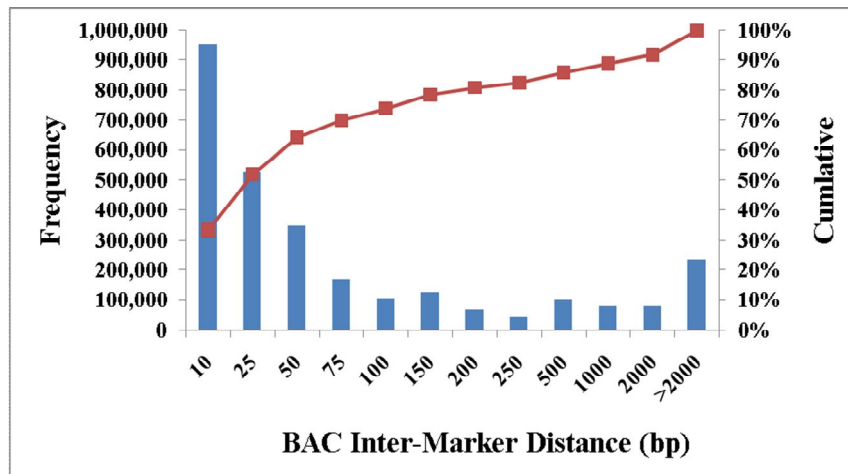


Figure 4.4. Distribution of BAC inter-marker distances for 2,867,766 non-redundant SNPs and indels. Inter-marker distances were only calculated within BACs. For ease of illustration, inter-marker distances >2000 bp were grouped.

We used a Fisher's exact test to determine if the scored polymorphisms mapped to a single locus or multiple loci (see MATERIALS AND METHODS). Of the 2.8 million non-redundant polymorphisms that were scored, the majority (65.1%) mapped to a single locus and of these 1,673,332 were SNPs. It was not surprising that 10.5% of the 2.86 million polymorphisms were scored as "paralogous," because it has been previously shown that at least one-third of an estimated 59,000 genes in the maize genome are duplicated (Blanc and Wolfe, 2004; Emrich et al., 2007; Messing et al., 2004). An additional 24.4% were scored as "potentially paralogous", because the ratio of reference to alternative alleles was in agreement with both the sequencing error rate (~1%) and expectations for paralogy (Max. Threshold= 0.01) (see "MATERIALS AND METHODS"). With the addition of base quality score as another metric, it should be possible to minimize the number of false positives within the potentially paralogous class that were due to sequencing errors. Additionally, if several recombinant inbred lines (RILs) from the NAM population are SBS

resequenced, we will be able to use identity-by-descent (IBD) between founders and RIL progeny to decrease the total rate of false polymorphism discovery as well as improve the accuracy of the genome assembly.

Genome variation and genetic diversity

Even though permissive alignment parameters (allowing 2 SNPs and ≤ 3 bp indels) were employed for SBS read mapping, a modest fraction of low-copy-number reads could not be mapped to the maize B73 reference genome sequence. Compared to the number of unmapped, low-copy-number B73 reads, an average of 8.67% low-copy-number reads in non-B73 founder lines were unmapped. This was most likely the consequence of high rates of divergence or insertion relative to the reference genome sequence (Figure 4.5). This estimate is in close agreement with observations based on the hybridization of gene-derived, 40 bp oligonucleotide (oligo) probes to B73 and Mo17 BAC libraries, which showed that 9.2% of hybridized probes were unique to Mo17 (Morgante et al., 2005). Based on the findings in our study, it is estimated that 25.7% of the low-copy-number space in the founders is highly dissimilar or inserted relative to the genome of B73 (Figure 4.5). It is possible that these unmapped SBS reads correspond to gene or pseudogene fragments that were copied and inserted by the *Helitron* family of transposable elements, a mechanism that has been implicated in the generation of extreme haplotype variability (e.g., presence/absence of genes) among maize inbred lines (Brunner et al., 2005; Fu and Dooner, 2002; Lai et al., 2005; Morgante et al., 2005).

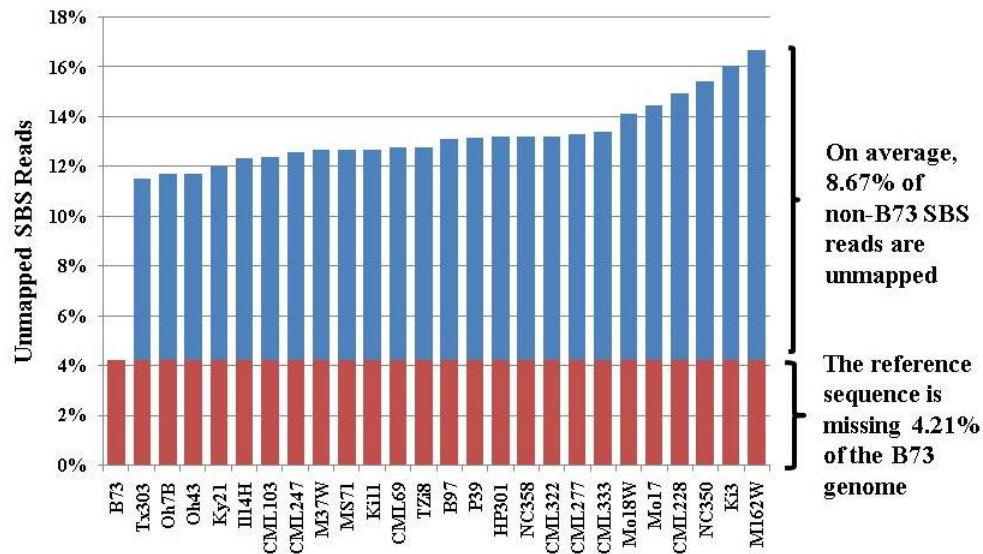


Figure 4.5. Distribution of unmapped low-copy-number SBS reads across the founder lines. Based on the observed number of unmapped B73 reads, 4.21% of the B73 genome was absent from the draft reference genome sequence. On average, 8.67% of non-B73 reads were unmapped when accounting for this missing B73 fraction. The non-B73 reads most likely correspond to sequences that are highly divergent or inserted relative to the B73 reference genome sequence. We estimated the 25.7% of the low-copy-number space in the founders is highly divergent or inserted relative to B73 by multiplying 8.67% by the corrected sample size for nucleotide diversity.

We measured the level of genetic diversity at the nucleotide level by calculating pairwise nucleotide diversity (π) (Nei, 1987). Nucleotide diversity was estimated to have an average value of 0.45% per BAC, but ranged nearly 3 orders of magnitude among BACs—from 0.0013 to 1.20%. This average is 55 to 68% lower than that of Sanger sequencing-based estimates for maize genes ($\pi=1-1.4\%$) (Tenaillon et al., 2001; Wright et al., 2005). The relatively low estimate of π based on Illumina SBS technology is primarily due to the less than expected number of singleton and doubleton SNPs in the dataset, which should theoretically account for 43% of SNPs (Buckler et al, in review) (Figure 4.6). These low-frequency SNPs are absent from the dataset, because only moderate to high frequency SNPs with

genotypic signal passed our stringently set thresholds. Machine learning algorithms will be used to improve the calling of singletons and doubletons (Matukumalli et al., 2006). In addition, our inability to align highly divergent haplotypes slightly lowered the estimates of π .

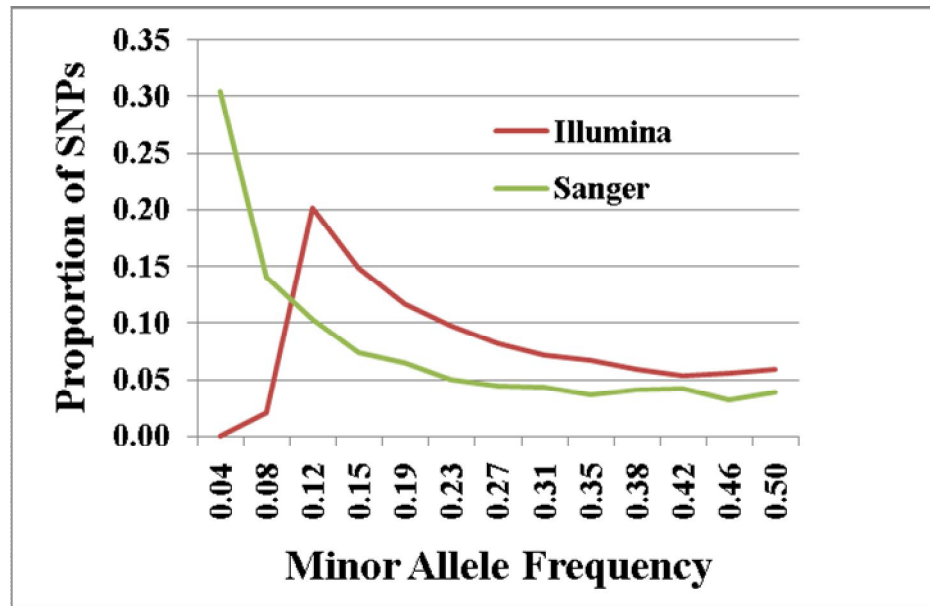


Figure 4.6. Minor allele frequency (MAF) distributions for SNPs discovered by Sanger and Illumina SBS sequencing technologies on 27 diverse maize lines. Singletons (0.04 MAF) and doubletons (0.08 MAF) account for 43% of SNPs discovered via capillary-based Sanger sequencing of amplicons from the 27 diverse lines (Buckler, et al., in review). Conversely, singletons and doubletons only represent 1.2% of SNPs discovered using SBS technology. These distributions are based on 3,641 Sanger-based SNPs (www.panzea.org) and 15,817 SBS-based SNPs that were scored on all 27 lines.

The presence of a non-random pattern of genome-wide nucleotide diversity was confirmed by first-order autocorrelation statistics (Figure 4.7). In general, the pattern of nucleotide diversity along the maize genome consisted of a large number of alternating “narrow peaks” and “wide valleys” (Figure 4.8). The 5% most diverse

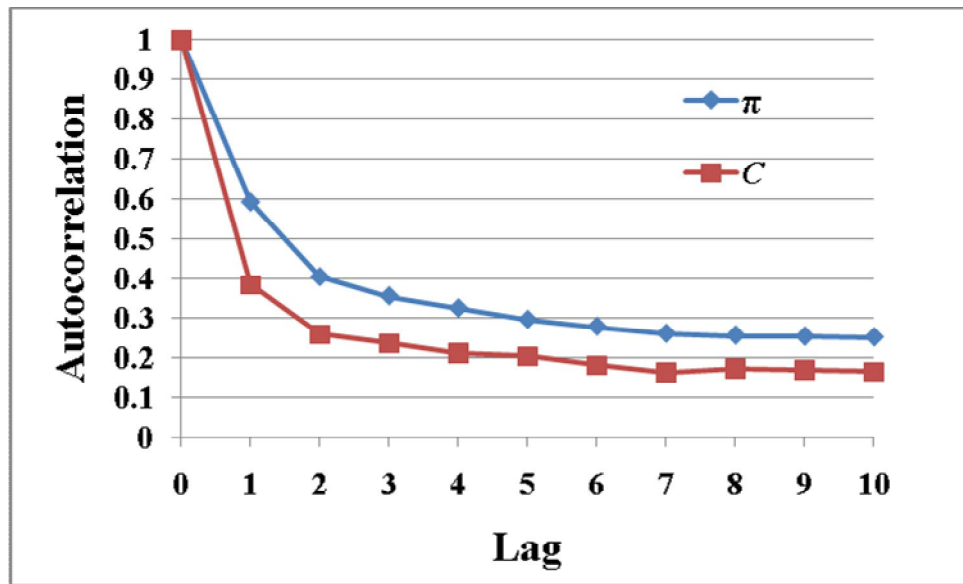


Figure 4.7. Autocorrelation plot of pairwise nucleotide diversity (π) and population-recombination (C) as a function of physical distance. There were not any instances of overlapping BAC sequences at lag 2, but there were at lag 1. There was a moderately high autocorrelation of 0.40 for π at lag 2 (Durbin-Watson test; $P < 0.0001$), which gradually decreased to a background level autocorrelation of 0.25. In the case of C , there was a modest autocorrelation of 0.30 at lag 2 (Durbin-Watson test; $P < 0.0001$). The background level autocorrelation for C was 0.17. The approximate physical distance between consecutive lags is 157 Kb (average BAC size).

BACs formed 123 multi-BAC peaks that had a mean size of ~340 kb. In contrast, the 5% least diverse BACs comprised 122 multi-BAC valleys that had a mean size of ~600 kb and an underlying distribution that was statistically significant from that of the multi-BAC peaks (Wilcoxon-Mann-Whitney test; $P < 0.0001$). Moreover, 21 (17.2%) of the multi-BAC valleys were >800 Kb in size, whereas all of the multi-BAC peaks were <800 Kb. Not all of the >800 Kb valleys were associated with centromeric or pericentromeric regions (Figure 4.8), which have been shown here and in previous work (Rafalski and Ananiev, 2009) to contain extensive stretches of low nucleotide diversity.

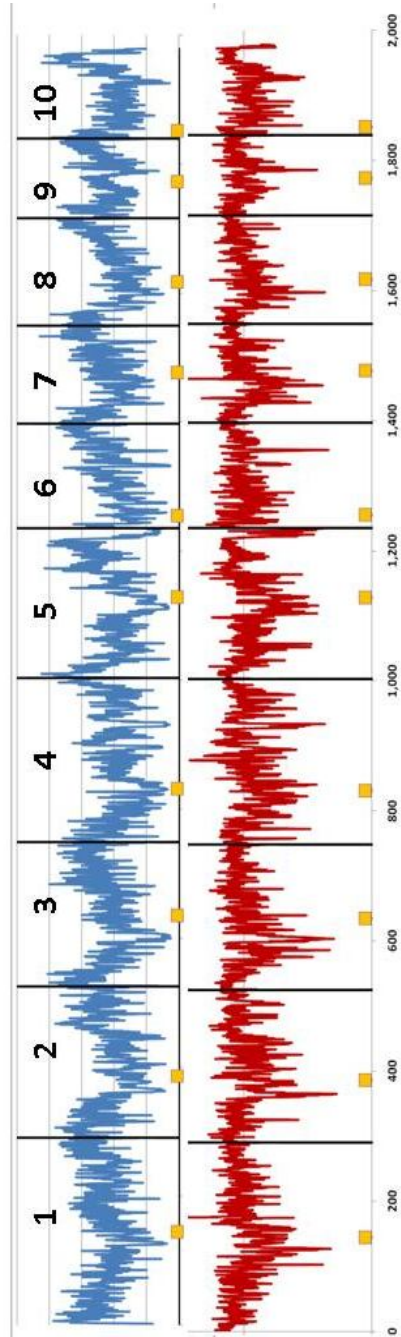


Figure 4.8. Genome-wide patterns of pairwise nucleotide diversity (π) and population-recombination (C). Pairwise nucleotide diversity (blue) and population-recombination (red) were plotted as a BAC average along each chromosome. Yellow squares denote the purported location of centromeres, and physical distance is measured in Mb. In general, there was a reduction in π and C near the centromeres, but π and C were also significantly reduced at other numerous locations throughout the genome.

Evolutionary forces appear to have shaped genetic diversity

Selection offers one explanation for the observed patterns of diversity across the maize genome. Two contrasting models predict an interaction between selection and recombination could explain these observed levels and patterns of nucleotide diversity. Under the selective-sweep model, diversity is lost as selection fixes advantageous alleles and nearby linked variants (Kaplan et al., 1989; Maynard-Smith and Haigh, 1974). Conversely, the background selection model postulates that diversity is lost as deleterious alleles are continuously purged (i.e., negative selection) from a population (Charlesworth et al., 1993; Hudson and Kaplan, 1995). Both models posit a loss of diversity at sites linked to a locus under selection, and that the reduction in polymorphism is most extreme in regions of low recombination. In support of the first hypothesis, detailed analysis of one of the diversity valleys on chromosome 10 suggests the effect of strong positive selection—a process that greatly reduced the diversity of 15 or more genes (Tian et al., 2009). In addition, the local recombination rate at this 1.1 Mb “sweep” was estimated to be 5-fold lower than the genome average. However, none of the other 20 valleys >800 Kb in size surrounded known major effect genes that were a target of strong positive selection during maize evolution, which implies that other uninvestigated traits have been the target of selection.

To gain further insight into the importance of selection, we investigated the influence of recombination on nucleotide diversity by estimating the population-recombination parameter C ($4N_e c$), which measures the rate of recombination in the history of a population (Hudson, 1987). Population-recombination was estimated to have an average value of 0.002 ± 0.0005 per site, but ranged 1350-fold among BACs. In general, this average is lower than previously reported estimates based on maize genes (Tenaillon et al., 2001; Tenaillon et al., 2002), which tend to be more recombinationally active than other genomic regions (Fu et al., 2002; Fu et al., 2001;

Yao et al., 2002). Similar to nucleotide diversity, an autocorrelated pattern of peaks and valleys was observed for C across the maize genome (Figures 4.7 and 4.8). A significant positive correlation ($R^2 = 0.48$; $P < 0.0001$) was detected between estimates of nucleotide diversity (π) and C (Figure 4.9), as hitchhiking and background selection models predict that recombination and SNP diversity are positively correlated (Charlesworth et al., 1993; Hudson and Kaplan, 1995; Kaplan et al., 1989; Maynard-Smith and Haigh, 1974). Although to explicitly detect the effects of hitchhiking and background selection, we will need to empirically test for signatures of selection (e.g. Tajima's D and HKA Tests) (Hudson et al., 1987; Tajima, 1989). Regardless of the evolutionary force, the strength of the correlation between π and C exceeds what has been previously reported in *Drosophila melanogaster* (Begun and Aquadro, 1992) and *Arabidopsis thaliana* (Kim et al., 2007).

In addition to examining C , a physical estimate of recombination (R) was determined by calculating the ratio of genetic (cM) and physical (Mb) distances for every pair of contiguous markers on the multi-population NAM genetic linkage map (McMullen et al., in review). A significant positive correlation ($R^2 = 0.24$; $P < 0.0001$) was detected between C (i.e., ancestral recombination) and R (i.e., recent recombination), which indicates that some rates and patterns of recombination in the genome have remained unchanged throughout the evolution of maize. In contrast to C , which is inversely related to the expected amount of linkage disequilibrium (LD) between segregating sites, R measures recombination rate per physical distance and thus is unaffected by population history, selection, and demography (e.g., population bottleneck, population size, etc.) (Pritchard and Przeworski, 2001; Tenailon et al., 2002). In that light, we detected a strong positive correlation ($R^2 = 0.32$; $P < 0.0001$) between R and π , which suggests that other factors such as genome structure have also affected the levels and patterns of nucleotide diversity in maize.

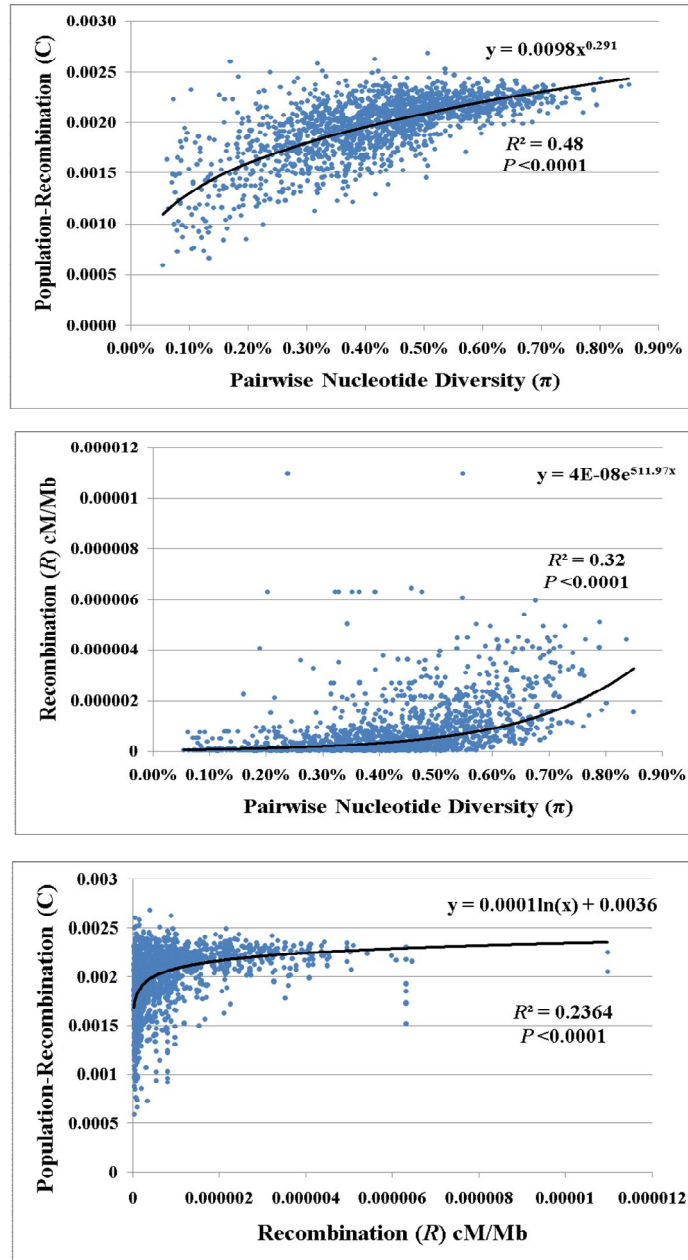


Figure 4.9. Correlations between two estimates of recombination (R and C) and between estimates of recombination and nucleotide diversity (π). These correlations were based on 1,686 BACs. Non-linear regression lines, coefficient of determination (R^2), and P values are given.

Impact of genome features on genetic diversity

In an attempt to identify which attributes of the genome impacted nucleotide

diversity, we partitioned the effect of recombination rate (R) from that of several sequence features using a stepwise regression analysis. Repeat content, CpG content, gene content, GC content, and distance to telomeres and centromeres were included in the analysis, because all have previously been shown to associate with recombination and/or mutation rate (Hellmann et al., 2005; Kong et al., 2002). We found that R and two other sequence features explained 33.7% of the total variance in nucleotide diversity (π) levels. The single best predictor of nucleotide diversity was R ($R^2 = 0.29$), followed by repeat content ($R^2 = 0.03$) and CpG content ($R^2 = 0.01$). Interestingly, the examined sequence features were found to be stronger predictors of R rather than of nucleotide diversity (Figure 4.10), which supports the claim that genome structure and composition underpin the recombination-rate gradient along maize chromosomes (Fengler et al., 2007; Schnable et al., 1998).

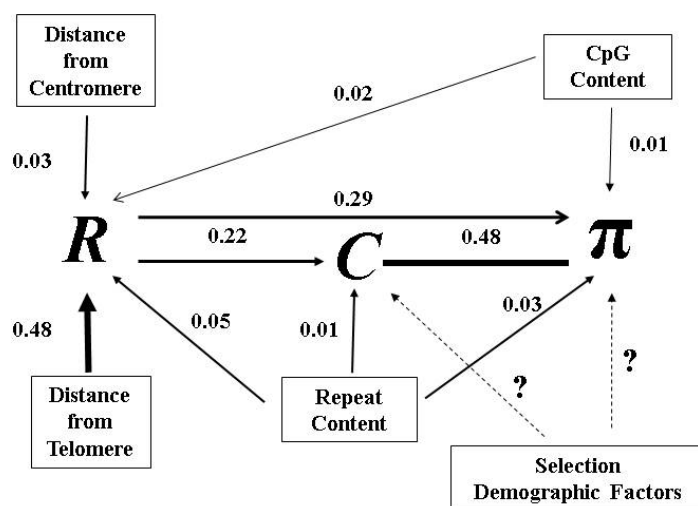


Figure 4.10. Plot of the correlation between sequence features and two estimates of recombination (R and C) as well as pairwise nucleotide diversity (π). The proportion of the total variance explained (partial R^2) by a statistically significant association is numerically denoted. Only sequence features that were significant in a multiple linear regression model using a stepwise procedure are included in the illustration. Gene content and GC content were significantly correlated with R , C , and π in a single linear regression model, but not when using multiple linear regression with a stepwise procedure. These correlations were based on a subset of 1,686 BACs.

Despite the inclusion of several sequence features, recombination remained the principal predictor of nucleotide diversity levels in maize, as was concluded in human (Hellmann et al., 2005). Recombination itself could be mutagenic, although strong evidence of this phenomenon in plants and other multi-cellular organisms is limited (Gaut et al., 2007). To investigate the likelihood of this phenomenon, we will need to estimate and compare the sequence divergence rates between maize and sorghum at these surveyed regions (Begun and Aquadro, 1992). It is unknown whether repeat content and CpG content truly associate with mutation rate, as it is possible that these two features associate with other factors that directly impact mutation rate such as epigenetic events, chromatin structure, or DNA replication rate (Hellmann et al., 2005; Jensen-Seaman et al., 2004; Walser et al., 2008). Taken together, these results imply that the interplay between selection and recombination, along with other demographic factors (e.g., selection bottleneck) rather than differential mutation rate has primarily caused diversity to vary substantially across the maize genome. Similar findings were reported in studies of *Drosophila* (Begun and Aquadro, 1992) and human (Hellmann et al., 2005).

Evidence for local selection in breeding populations

To measure the amount of genetic variation within and between temperate and Tropical or Semitropical (TS) subpopulations, F_{ST} was calculated for each BAC. F_{ST} measures the proportion of the total genetic variance in a subpopulation as a fraction of the total genetic variance (Weir and Cockerham, 1984). The 27 diverse lines were separated into temperate or TS subpopulations according to the findings of a previous model-based clustering analysis (Liu et al., 2003). The temperate group consisted of seven Non Stiff Stalk (NSS) lines, including a Stiff Stalk (SS; B73) line and three specialty lines (HP301 popcorn; Il14H and P39 sweet corn) that have a large

contribution from Northern Flint germplasm. The TS group consisted of 16 maize lines that are more representative of lowland than highland tropical germplasm.

F_{ST} between groups was 0.038 ± 0.045 —a low level of differentiation—with F_{ST} for each BAC ranging from -0.06 to 0.32. Similarly, Liu et al. (2003) showed a low level of differentiation between TS and NSS ($F_{ST}= 0.06$), but their study showed increased differentiation between TS and SS ($F_{ST}= 0.47$) as well as the specialty lines and TS ($F_{ST}= 0.52$ -0.58). In general, moderate to high frequency SNPs were shared between temperate and TS groups, although there were 157 BACs with an F_{ST} value greater than 0.18 ($P= 8.2 \times 10^{-4}$). It is possible that local selection has resulted in these extreme F_{ST} values, as large F_{ST} values have been used to identify candidate loci that have undergone recent positive selection in human populations (Hinds et al., 2005). If this is indeed the case, one would suspect that these regions contain genes underlying adaptive traits such as flowering time, photoperiod sensitivity, temperature adaptation, or stress tolerance. We need to further investigate if adaptive trait quantitative trait loci (QTL) identified in the NAM population co-localize with these regions purportedly involved in local adaptation.

Prospects for GWAS in diverse maize

Linkage disequilibrium (LD), the non-random association of alleles at different loci (Hedrick, 1987), was analyzed at genome-wide resolution across the 27 diverse lines. In concordance with previously reported LD estimates in maize (Remington et al., 2001; Whitt et al., 2002), on average, LD decays to nominal levels (mean $r^2 < 0.10$) within 2,300 bp and reached 50% of its starting value at about 100-200 bp (Figure 4.11). The extent of LD also varied greatly across the genome, as it is essentially the inverse of C (Pritchard and Przeworski, 2001). In general, the rate of LD decay was rapid in regions of high recombination and diversity, and vice versa.

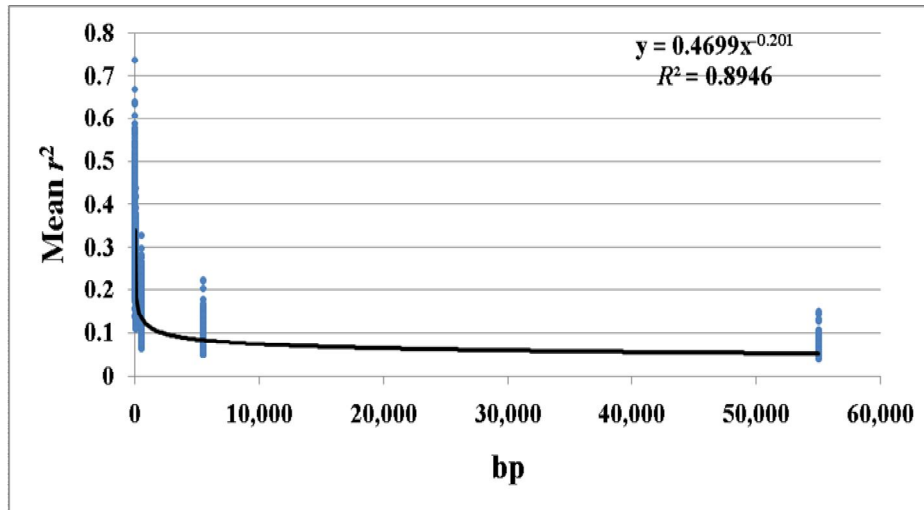


Figure 4.11. The average decay of LD across the length of a BAC. Mean r^2 was plotted as a function of average distance between SNPs, and mean r^2 was calculated at distances of 1-10 bp, 11-100 bp, 101-1,000 bp, 1,001-10,000 bp, and 10,001-100,000 bp within a BAC. The LD plot was based on 1,247 BACs that had at least 50 SNPs at a distance of 1-10 bp. Non-linear regression lines, coefficient of determination (R^2), and P value are given.

A major impetus for this work was to discover markers for genome-wide association studies (GWAS) in diverse maize. With the current dataset of 1.67 million single-locus SNPs, we estimated that only ~50% of the SNPs were in high LD ($r^2 > 0.8$) with at least one other SNP in the dataset (data not shown). Therefore, to obtain maximum power for GWAS using the maize nested association mapping (NAM) population, we will need to construct a more complete HapMap with contiguous coverage to ensure that almost every polymorphism is in high LD ($r^2 \geq 0.8$) with a robustly scored SNP (Wang et al., 2005). Given the rapid rate of LD decay in a highly diverse genome, we estimate that we will need to identify a total of 10-30 million SNPs for powerful GWAS in diverse maize. If this goal is to be achieved, it will be necessary to resequence nearly the entire genome of each founder line.

REFERENCES

- The Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69-87.
- Begun, D.J., and C.F. Aquadro. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356:519-520.
- Bennett, S. 2004. Solexa Ltd. *Pharmacogenomics* 5:433-438.
- Blanc, G., and K.H. Wolfe. 2004. Widespread Paleopolyploidy in Model Plant Species Inferred from Age Distributions of Duplicate Genes. *Plant Cell* 16:1667-1678.
- Brunner, S., K. Fengler, M. Morgante, S. Tingey, and A. Rafalski. 2005. Evolution of DNA sequence nonhomologies among maize inbreds. *Plant Cell* 17:343-60.
- Buckler, E.S., B.S. Gaut, and M.D. McMullen. 2006. Molecular and functional diversity of maize. *Curr. Opin. Plant Biol.* 9:172-176.
- Charlesworth, B., M.T. Morgan, and D. Charlesworth. 1993. The Effect of Deleterious Mutations on Neutral Molecular Variation. *Genetics* 134:1289-1303.
- Emberton, J., J. Ma, Y. Yuan, P. SanMiguel, and J.L. Bennetzen. 2005. Gene enrichment in maize with hypomethylated partial restriction (HMPR) libraries. *Genome Res.* 15:1441-1446.
- Emrich, S.J., L. Li, T.-J. Wen, M.D. Yandea-Nelson, Y. Fu, L. Guo, H.-H. Chou, S. Aluru, D.A. Ashlock, and P.S. Schnable. 2007. Nearly Identical Paralogs: Implications for Maize (*Zea mays* L.) Genome Evolution. *Genetics* 175:429-439.
- Fengler, K., S.M. Allen, B. Li, and A. Rafalski. 2007. Distribution of Genes, Recombination, and Repetitive Elements in the Maize Genome. *Crop Sci.* 47:S-83-95.
- Fu, H., and H.K. Dooner. 2002. Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci.* 99:9573-9578.
- Fu, H., Z. Zheng, and H.K. Dooner. 2002. Recombination rates between adjacent genic and retrotransposon regions in maize vary by 2 orders of magnitude. *Proc. Natl. Acad. Sci.* 99:1082-1087.

- Fu, H., W. Park, X. Yan, Z. Zheng, B. Shen, and H.K. Dooner. 2001. The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. *Proc. Natl. Acad. Sci.* 98:8903-8908.
- Gaut, B.S., S.I. Wright, C. Rizzon, J. Dvorak, and L.K. Anderson. 2007. Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.* 8:77-84.
- Hedrick, P.W. 1987. Gametic Disequilibrium Measures: Proceed With Caution. *Genetics* 117:331-341.
- Hellmann, I., K. Prufer, H. Ji, M.C. Zody, S. Paabo, and S.E. Ptak. 2005. Why do human diversity levels vary at a megabase scale? *Genome Res.* 15:1222-1231.
- Hinds, D.A., L.L. Stuve, G.B. Nilsen, E. Halperin, E. Eskin, D.G. Ballinger, K.A. Frazer, and D.R. Cox. 2005. Whole-Genome Patterns of Common DNA Variation in Three Human Populations. *Science* 307:1072-1079.
- Holland, J.B. 2007. Genetic architecture of complex traits in plants. *Curr. Opin. Plant Bio.* 10:156-161.
- Hudson, R.R. 1987. Estimating the recombination parameter of a finite population model without selection. *Genet. Res.* 50:245-251.
- Hudson, R.R., and N.L. Kaplan. 1995. Deleterious Background Selection With Recombination. *Genetics* 141:1605-1617.
- Hudson, R.R., M. Kreitman, and M. Aguade. 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* 116:153-159.
- Jensen-Seaman, M.I., T.S. Furey, B.A. Payseur, Y. Lu, K.M. Roskin, C.-F. Chen, M.A. Thomas, D. Haussler, and H.J. Jacob. 2004. Comparative Recombination Rates in the Rat, Mouse, and Human Genomes. *Genome Res.* 14:528-538.
- Kaplan, N.L., R.R. Hudson, and C.H. Langley. 1989. The "Hitchhiking Effect" Revisited. *Genetics* 123:887-899.
- Kim, S., V. Plagnol, T.T. Hu, C. Toomajian, R.M. Clark, S. Ossowski, J.R. Ecker, D. Weigel, and M. Nordborg. 2007. Recombination and linkage disequilibrium in *Arabidopsis thaliana*. *Nat Genet.* 39:1151-1155.
- Kong, A., D.F. Gudbjartsson, J. Sainz, G.M. Jonsdottir, S.A. Gudjonsson, B. Richardsson, S. Sigurdardottir, J. Barnard, B. Hallbeck, G. Masson, A. Shlien, S.T. Palsson, M.L. Frigge, T.E. Thorgeirsson, J.R. Gulcher, and K. Stefansson. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.*

31:241-247.

- Lai, J., Y. Li, J. Messing, and H.K. Dooner. 2005. Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci.* 102:9068-9073.
- Liu, K., M. Goodman, S. Muse, J.S. Smith, E. Buckler, and J. Doebley. 2003. Genetic Structure and Diversity Among Maize Inbred Lines as Inferred From DNA Microsatellites. *Genetics* 165:2117-2128.
- Matukumalli, L., J. Grefenstette, D. Hyten, I.-Y. Choi, P. Cregan, and C. Van Tassell. 2006. Application of machine learning in SNP discovery. *BMC Bioinformatics* 7:4.
- Maynard-Smith, J., and J. Haigh. 1974. The hitch-hiking effect of a favorable gene. *Genet. Res.* 23:23-35.
- Messing, J., A.K. Bharti, W.M. Karlowski, H. Gundlach, H.R. Kim, Y. Yu, F. Wei, G. Fuks, C.A. Soderlund, K.F.X. Mayer, and R.A. Wing. 2004. Sequence composition and genome organization of maize. *Proc. Natl. Acad. Sci.* 101:14349-14354.
- Morgante, M., S. Brunner, G. Pea, K. Fengler, A. Zuccolo, and A. Rafalski. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.* 37:997-1002.
- Nei, M. 1987. *Molecular Evolutionary Genetics*. Columbia Univ. Press, New York.
- Pritchard, J.K., and M. Przeworski. 2001. Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* 69:1-14.
- Rabinowicz, P.D., L.E. Palmer, B.P. May, M.T. Hemann, S.W. Lowe, W.R. McCombie, and R.A. Martienssen. 2003. Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res.* 13:2658-2664.
- Rabinowicz, P.D., R. Citek, M.A. Budiman, A. Nunberg, J.A. Bedell, N. Lakey, A.L. O'Shaughnessy, L.U. Nascimento, W.R. McCombie, and R.A. Martienssen. 2005. Differential methylation of genes and repeats in land plants. *Genome Res.* 15:1431-1440.
- Rafalski, A., and E. Ananiev. 2009. Genetic Diversity, Linkage Disequilibrium and Association Mapping, p. 201-219, *In* J. L. Bennetzen and S. Hake, eds. *Handbook of Maize Genetics and Genomics*.

- Remington, D.L., J.M. Thornsberry, Y. Matsuoka, L.M. Wilson, S.R. Whitt, J. Doebley, S. Kresovich, M.M. Goodman, and E.S. Buckler. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl. Acad. Sci.* 98:11479-11484.
- Salvi, S., and R. Tuberosa. 2005. To clone or not to clone plant QTLs: present and future challenges. *Trends Plant Sci.* 10:297-304.
- Schnable, P.S., A.-P. Hsia, and B.J. Nikolau. 1998. Genetic recombination in plants. *Curr. Opin. Plant Bio.* 1:123-129.
- Tajima, F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. *Genetics* 123:585-595.
- Tenaillon, M.I., M.C. Sawkins, A.D. Long, R.L. Gaut, J.F. Doebley, and B.S. Gaut. 2001. Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). *Proc. Natl. Acad. Sci.* 98:9161-9166.
- Tenaillon, M.I., M.C. Sawkins, L.K. Anderson, S.M. Stack, J. Doebley, and B.S. Gaut. 2002. Patterns of Diversity and Recombination Along Chromosome 1 of Maize (*Zea mays* ssp. *mays* L.). *Genetics* 162:1401-1413.
- Tian, F., N.M. Stevens, and E.S. Buckler. 2009. Tracking footprints of maize domestication and evidence for a massive selective sweep on chromosome 10. *Proc. Natl. Acad. Sci.* In press.
- Walser, J.-C., L. Ponger, and A.V. Furano. 2008. CpG dinucleotides and the mutation rate of non-CpG DNA. *Genome Res.* 18:1403-1414.
- Wang, W.Y.S., B.J. Barratt, D.G. Clayton, and J.A. Todd. 2005. Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet* 6:109-118.
- Weir, B.S., and C. Cockerham. 1984. Estimating F-Statistics for the Analysis of Population Structure. *Evolution* 38:1358-1370.
- Weir, B.S. 1996. *Genetic Data Analysis II*. Sinauer, Sunderland, MA.
- Whitt, S.R., L.M. Wilson, M.I. Tenaillon, B.S. Gaut, and E.S. Buckler. 2002. Genetic diversity and selection in the maize starch pathway. *Proc. Natl. Acad. Sci.* 99:12959-12962.
- Wright, S.I., I.V. Bi, S.G. Schroeder, M. Yamasaki, J.F. Doebley, M.D. McMullen, and B.S. Gaut. 2005. The effects of artificial selection on the maize genome. *Science* 308:1310-1314.

- Yao, H., Q. Zhou, J. Li, H. Smith, M. Yandeu, B.J. Nikolau, and P.S. Schnable. 2002. Molecular characterization of meiotic recombination across the 140-kb multigenic *a1-sh2* interval of maize. *Proc. Natl. Acad. Sci.* 99:6157-6162.
- Yu, J., and E.S. Buckler. 2006. Genetic association mapping and genome organization of maize. *Curr. Opin. Biotech.* 17:155-160.
- Yu, J., J.B. Holland, M.D. McMullen, and E.S. Buckler. 2008. Genetic design and statistical power of nested association mapping in maize. *Genetics* 178:539-551.
- Yuan, Y., P.J. SanMiguel, and J.L. Bennetzen. 2002. Methylation-spanning linker libraries link gene-rich regions and identify epigenetic boundaries in *Zea mays*. *Genome Res.* 12:1345-1349.